Project Title: Parallel and Distributed Computational Models for Scientific Applications on Cell Processor Clusters

IBM R&D Contact: Roland Seiffert - Lead Architect, Linux on Cell BE

Principal Researcher: Salvador Abreu - spa@di.uevora.pt (CENTRIA)
Contact Person: Paulo Lopes - pal@di.fct.unl.pt (CITI)

1. Introduction and Scope

CENTRIA and CITI are research centres hosted at *Departamento de Informática* of the *Faculdade de Ciências e Tecnologia* of *Universidade Nova de Lisboa* (DI-FCT/UNL). These centres incorporate researchers from other universities, in particular from *Universidade de Évora* (UE) but also from *Faculdade de Ciências* of *Universidade de Lisboa* (FC/UL) and *Universidade de Coimbra* (UC).

Researchers in both centres have developed pioneering work in Portuguese CS and are teaching one of best known and the oldest computer science degree in Portugal, and currently teach a 3-year Bologna-style 1st cycle (*Licenciatura em Engenharia Informática*) and two (Master) 2nd cycles: *Mestrado em Engenharia Informatica* and the European Master in Computational Logic (EMCL: CENTRIA, jointly with T.U. Dresden in Germany, U.P. Madrid in Spain, F.U. Bolzano in Italy and T.U. Wien in Austria). Students from these programmes are cooperating in many CENTRIA and CITI projects and are thus expected users of the resources provided by this proposal. *Universidade de Évora* carries similar educational curricula.

The present proposal will support the ongoing trend (and need) of applying fundamental CS research to other scientific areas. Granted equipment will further enhance the computational infrastructure that supports our ongoing research, which includes joint work with teams in other scientific areas.

One of the research directions of the present proposal – Bioinformatics and Constraint Programming – just had a key researcher (Ludwig Krippahl) distinguished with the Portuguese IBM scientific award.

2. Research Directions

The research groups of CENTRIA and CITI are involved in several projects which will benefit from the resources which will be allocated as a result of the present proposal. In the scope of this SUR application, we specifically target the following areas[1]:

- Computer Science Concepts & Technologies
  - Constraint Programming
  - Parallel and Distributed Computing (including Cluster & Grid)

    In particular, this includes Constraint Execution in multi-Cell clusters
- Application areas
  - Bioinformatics and Life Sciences
  - Earth, Sea and Space research

For the current SUR proposal the main focus is the development of techniques and tools which may effectively draw on the potential of hierarchical parallel computing infrastructures, which goes from heterogeneous multicore processors such as the Cell to cluster systems. This goal is achieved by exploiting computational models well suited to the architecture.

1 A more extensive description of CENTRIA and CITI research areas can be found in the Appendix under "Collaborative Research".

3. SUR-focused Collaborative Research

1 Constraint Satisfaction Problems

Several application problems may be described as CSPs (constraint satisfaction problems). The ability to solve such sets of constraints relies on heuristics which may make effective use of distributed algorithms.

CENTRIA, CITI and their associated partners are jointly developing distributed CSP algorithms and implementations which are designed to be well-suited to several types of parallel architectures, including NUMA-style shared memory multiprocessors, general-purpose clusters, and grids.

2 Short and medium term projects (ongoing & up to 2 years)

An MSc. student from *Universidade de Évora* is presently at IBM's Böblingen research centre, developing his thesis on the use of Cell SPEs working in parallel as partial hybrid constraint solver engines. This collaboration is expected to intensify through additional internships at IBM's research facility. This line of work is expected to have repercussions on other areas which rely on constraint programming tools and techniques, such is the case with Bioinformatics, for instance.

A project centered on the parallelisation of algorithms used to process data obtained from synchrotron radiation X-ray microtomography (regarding metal matrix composite characterization, emphasis being placed in the case of Al-based functionally graded materials) has just been completed: an MSc student parallelised the Vignoles algorithm, reaching a speedup factor of 8.9 when processing a 100 × 100 × 100 voxel dataset from an Al/SiCp FGMMC tomography on a small cluster with 10 CPUs. For this application, where processing power is required both for raw computing and visualization facilities, we wish to investigate the adequacy of the Cell processor cluster and compare it to other common off-the-shelf clusters.

Another application that could take profit of the capabilities of the Cell processor is a simulation of pollutant dispersion that is being developed in cooperation by CITI and a researcher of the Environmental Sciences Department. This work is based in novel methods of simulation that were the focus of a 2003 PhD thesis within the Environmental Sciences Department of FCT/UNL and Delft University in Holland; it is centered on numerical methods for advection-diffusion simulation and its main idea is the direct relation between pollutant particle displacement moments and truncation errors. This relation raised the theoretical foundations to create a new family of numerical methods, called DisPar. Several variants of this algorithm were developed and tested in a heterogeneous and non-dedicated PC Cluster. To do so, a new partitioning method has been developed, AORDA. The application, Scalable DisPar, was implemented with the Microsoft .NET environment and tested on the river Tagus Estuary, near Lisbon (Portugal). Since 2005, and using a cluster donnated by IBM in a previous SUR grant, CITI researchers have been working in alternative implementations of the simulator. In particular, we are interested in seeing how the strategies of work partitioning and dynamic load balance that proved to work in a NOW (Network of Workstations) created by harvesting CPU cycles from student labs' PCs, are suitable to dedicated parallel processing platforms like computational clusters and grids. In order to obtain this insight, the software was ported to the IBM Linux cluster, using the MONO implementation of .NET. First results suggest that alternative approaches are needed; we are also in the process of rewriting the application to use the MPI message-passing library, which, due to the change of communication paradigm, requires extensive rewriting of the code.

3 Long term projects (3+ years)

CENTRIA has been developing a number of artificial intelligent techniques that should be applied to data collected by collaborating centres in this proposal and with their assistance. Constraint programming has been successfully applied in the area of structural bioinformatics, not only to predict 3D structure of proteins (from constraints imposed by NMR data) as well as protein interaction (docking). These applications also require complementary techniques of machine learning and simulation, require addressing gene expression and several sequencing applications (homology modelling in proteins, but also DNA and RNA). Additionally, general knowledge representation techniques have been developed with a view on Semantic

Web application that will allow much more powerful intelligent access to bioinformatics data accessible in Internet servers and the integration of their services.

CITI has been pursuing some research areas which are relevant for this proposal, focused around the PLM and PDPS groups. The PLM group has been working on programming models, languages, and algorithms, applied to the area of systems biology and algorithms for data mining and searching in very large data sets. As a promising direction, there is ongoing work on the specification and verification of space-time properties of biological systems, and also on the development of spatial databases supporting queries based on sophisticated notions of distances between entities. The PDPS group has been working on parallel and distributed computing models and algorithms to support complex problem solving, and techniques enabling large-scale simulations with massive data archiving and parallel searching (as required in bioinformatics or natural language processing), have been developed, and are being improved.

We shall also, as a secondary goal (but important to IBM's visibility), contribute towards a FCT/UNL Campus Grid, which will be important to allow remote collaboration with other scientists accessing Bioinformatics Grids.

CENTRIA and CITI, together with other participating centres[2] – CEG, CIGA, CQFB and CREM – will collaborate on:

- Simulation of Complex Systems – In complement to direct observations and experimentation, the behaviour of complex systems (such as metabolic pathways in gene expression) can be studied through computer simulations, requiring the use of appropriate modelling and programming techniques. We intend to extend a technique developed in CENTRIA to incorporate differential equations in constraint programming, so as to guarantee safe bounds in unknown parameters, and include proper management of uncertainty and preferences in the models. Other more abstract modelling techniques (e.g. graph models, discrete simulation) will be explored and applied to bioinformatics systems.

- Algorithms for Bioinformatics – Constraint Programming has already been applied in bioinformatics for the determination of 3D structure of proteins, and their interactions; existing techniques will be complemented with advances in search and treatment of homologies, which require not only the integration of known algorithms based on sequences provided by bioinformatics servers, but also new algorithms that include additional information, such as secondary structures, and inferred graph properties of the representations.

- Geological Models – Adaptation and further development of constraint processing to continuous domains in order to use it in geological and geophysical applications. This will interface with research on new geological models and their validation, focusing on the following topics: to exploit integration of information with geological uncertainty, at different scales and resolutions, with 3D modelling and visualizations; building 3D models and support partially automated interpretation of geological information in depth; application of the above 3D models to urban problems, such as geological hazard risk events, contaminant dispersion, and erosion.

4. Expected Contribution from IBM

In order to support the various initiatives described herein, the proponents would benefit from having their shared computing resources extended; the projects mentioned herein will particularly benefit from the availability of a Cell-based cluster, as well as from a shared storage system which would ensure a high-performance, highly available, data storage infrastructure. The following configuration would constitute a useful starting point for a Cell-based cluster system:

2 See further down on this document for the list of participating centres and institutions.

- 1 BladeCenter Chassis H with power supplies and a GbE switch
- 1 JS21 blade server for infrastructural support (TFTP, NFS, etc.)
- 3 QS21 blade servers (the upcoming Cell blades)

This system will form the basis for further expansion, as the projects grow.

5. Benefits for IBM

Benefits to IBM are manifold:

- Research outcome is expected to prove the adequacy of the IBM granted hardware and software to efficiently tackle the computational problems within the targeted research domains.

- Greater exposure to IBM's software portfolio: both HPC-specific, to be used in research activities, or more "general purpose" (such as WebSphere, DB2, Rational, *etc*.) to be used in the development of applications and portals for storing and disseminating project information, bookkeeping activities, communication among members of the different user groups, *etc*.

- Increased visibility in the community

    - At the researcher and graduate levels, both at the national level, across the participant universities and research organisations, as well as worldwide, through both the European Master in Computational Logic (EMCL) and PhD students that come from all over the world, postdoc scholarships, and multiple ongoing international projects. In particular, there have been EMCL studentes from already over 20 different nationalities who stay one year at UNL and another at one of the other participating universities, thereby extending the immediate visibility of the awarded resources.

    - At the undergraduate level, as a new, promising, state-of-the-art technology will be made available for student experimentation, and will allows us to complement already ongoing p-Series and z-Series IBM lectured seminars, regularly scheduled along with Computer Architecture, Operating System, and Computer Systems Performance classes at DI-FCT/UNL. Similar benefits would be visible at U. of Evora, as graduate students will rely on the proposed platform.

    - Dissemination of IBM's role in supporting research efforts for the scientific community in general, and portuguese scientific community in particular.

    - Shared use of these resources, in particular as is the case for common projects carried out with researchers from other institutions, will be reflected in the published results. These include several ongoing collaborations, with partners in the European Union, the US and China.

Appendix

I. CENTRIA

The *Centro de Inteligência Artificial* (CENTRIA) is an institutionalised research centre of *Universidade Nova de Lisboa* (UNL), funded by the Ministry of Science and Higher Education, and comprising around 50 people, 26 of which hold a PhD degree and the remainder being research students. The centre is hosted by the *Departamento de Informática* of the *Faculdade de Ciências e Tecnologia* of UNL (DI-FCT/UNL) and includes researchers from other universities, including in particular *Universidade de Évora* (UE). CENTRIA is the first national AI centre, and its senior researchers are internationally recognized. It's main areas of activity are:

- Knowledge Representation and Reasoning
- Logic Programming
- Natural Language
- Constraints and Soft Computing
- Machine Learning
- Intelligent Information Systems

CENTRIA has been evaluated three times by an independent international committee appointed by the Ministry and consistently labelled "Very Good".

II. CITI

The *Centro de Investigação em Informática e Tecnologias da Informação* (CITI) is a research center partially funded by the Portuguese National Science Foundation (*Fundação para a Ciência e Tecnologia*, FCT) and by the DI-FCT/UNL, where it is located since its foundation in 1997. It's main research areas are:

- Programming Languages and Software Engineering
    - Programming Languages and Models (PLM)
    - Software Engineering
- Parallel and Distributed Computing Systems
    - Parallel and Distributed Processing Systems (PDPS)
    - Large Scale Distributed Computing Systems
- Computer Graphics, Multimedia and Human-Computer Interaction
    - Computer Graphics (CG)
    - Media Processing Visualization and Interaction (MPVI)
    - Human Language Technologies

The objective of CITI is to promote basic and applied research in Computer Science and Informatics. CITI research team is currently composed by more that 50 researchers, of which around 40 hold a PhD degree, thus registering an impressive growth in specialized human resources from the initial core of 8 PhDs back in 1997. While the vast majority of its members are FCT/UNL faculty, CITI has also acted as an attraction pole for researchers from other universities. The centre has been rated "very good" by the last evaluation organized by an independent international committee appointed by the funding entity, FCT.

III. Collaborative Research

The application of the research techniques and results to real problems has been a major objective of both CENTRIA and CITI, which aim to make the results of research available to the scientific community. The opportunity of developing applications in distinct fields, possibly taking into account new, advanced computational models is clearly well integrated in the strategic objectives of both centres.

1. Earth, Sea and Space Sciences

The climate and environmental models developed in the Oceanographic Institute (*Instituto Oceanográfico*, IO - FC/UL) urge the development of sophisticated and reliable monitoring and analysing techniques. Satellite remote sensing plays an important role in this issue.

The tools and methods under development in CENTRIA and the computing models and environments in CITI will help to achieve the above objectives by its application to automatic search and classification of relevant sea surface oceanographic structures present in the IO satellite data archive. From this joint effort a better interpretation of satellite images of the ocean and coastal zone and the development of advanced remote sensing monitoring programs will be expected. It is also predictable a contribution to a further understanding of the biology-physics interaction processes, providing a scientific basis for coastal zone management. CENTRIA and IO have been collaborating in the exploration of machine learning (both supervised and unsupervised) techniques to the automatic identification and analysis of patterns of oceanic mesoscale phenomena in the Iberian Coastal Ocean derived from satellite imagery. This collaboration is formalized in the ongoing GRICES/ESA project "RENA" (PDCTE/CTA/49945/2003) and in a new R&D joint project proposal recently submitted to FCT: 'Learning Spatio-Temporal Oceanographic Patterns' (PTDC/EIA/68183/2006). Collaboration is also ongoing with Universidade de Évora's Geophysics research centre, CGE/UE, on several fronts namely models for seismic tomography, for the modeling and visual rendering of the seismic wave propagation.

CITI will bring in three research groups – PDPS, CG and MPVI – to promote the following main topics:

- Parallel algorithms for computation and data intensive applications and distributed problem-solving environments for Earth, Space and Sea Sciences.

- Distributed monitoring and control for collaborative virtual laboratories, for application observation and computational steering of distributed simulations; for large-scale distributed sensor networks for remote data acquisition and monitoring.

- Grid computing environments to enable access to distributed computational resources and information repositories, with high-performance, collaborative user interaction; and access to large remote data repositories.

- Models to simulate, process and visualize complex data; computer graphics algorithms for 3D.

Researchers hosted by CENTRIA, CEG, CIGA and CITI work at the campus of FCT/UNL, where these centres and computing laboratories are located. The researchers hosted by IO work in the campus of FC/UL where the laboratories for Physical Oceanography, Marine Zoology and Marine Botany, an Instrument Calibration Laboratory and a Space Oceanography Centre are located.

2. Bioinformatics

Bioinformatics is an area of great importance for CENTRIA in that many Artificial Intelligence techniques (knowledge representation in general, and more specifically search and constrained optimisation, automated machine learning, complex systems simulation) have a direct application to the wealth of data that life sciences (namely biochemistry, biology and medicine) have been accumulating in recent years. All the more

so as, in addition to the data obtained from Life Sciences centres were active ongoing collaboration exists. Namely CQFB (FCT-UNL) and CREM. CQFB (FCT-UNL) is a Research Centre that together with CQUP formed the largest Associated Laboratory (REQUIMTE) with a strong impact in the area of Green Chemistry (Sustainable Chemistry). Much of this data is publicly accessible in the Internet through different bioinformatics servers and services. The group of the Marie Curie Chair at ICAM (UÉvora) is interested in studying the effect of DNA polymorphism on protein structure variability.

The research work already developed in the area of structural bioinformatics has in fact been acknowledged by the panel of external reviewers of CENTRIA that explicitly proposed funding this particular domain within the Centre, and we intend to develop it further, as already stated in the "SUR-focused Research", and further complemented with:

- Knowledge Representation and Machine Learning – Search for relevant bioinformatics information must take into account a variety of sources, from the classification of organisms and structures to phylogenetic evolution. Much information maintained in the Internet hides relationships not completely understood, or even suspected. Research already carried on in CENTRIA on data mining (in protein databanks) and knowledge representation should be extended and tested in different applications (determination of active sites in structures, search for different kinds of homologies, such as domains or secondary structures), and,

- Access to large (Bioinformatics) data banks – In contrast with other scientific areas, a large number of Bioinformatics data banks is publicly and freely available in the Internet. However, efficient access to these data banks is largely dependent on the development of the Semantic Web, whereby access to structured meta-data may allow automating the access to bioinformatics data. Work in logic based knowledge representation and reasoning and its application to the Semantic Web will be further developed with a view on Bioinformatics applications, namely the integration of different data sources for problem solving (*e.g.* structural bioinformatics).

CITI's role in this task will be to bring in four research groups – PLM, PDPS, CG and MPVI – to promote the already mentioned topics, under "SUR-focused Collaborative Research", plus the following ones:

- Concurrency in Biological Systems: Specification and verification of space-time properties of complex systems, for modelling biological systems.
- Parallelism in Bioinformatics Computations: Parallel and distributed computing to allow efficient simulations and access to massive data storage archives.
- Grid Computing Models and Environments: Grid Computing to support Bioinformatics.

3. COGNOMA

CENTRIA work in intelligent man-machine Interfaces is inspired by the text relative to "Knowledge Systems, Cognition and Learning" included in the European Community document (Dossier 2005/0185 - CNS). This text refers the need for methods and techniques to acquire and interpret, represent and personalize, navigate and retrieve, share and deliver knowledge, recognizing the semantic relationships in content for use by humans and machines; artificial systems that perceive, interpret and evaluate information, and that can cooperate, act autonomously and learn; theories and experiments that move beyond incremental advances benefiting from insights into natural cognition, in particular learning and memory, also for the purpose of advancing systems for human learning. This emphasis can be understood via the notion of "cognoma", by analogy with "genoma". *Cognoma* is a concept with three components, meaning the cognition in the human being, in the machine, and in the interface between both. In 2006, CENTRIA and GECAD R&D units presented a joint proposal to the Portuguese Science and Technology Foundation (FCT) towards the constitution of an Associated Laboratory, also titled COGNOMA, which can be consulted for details.

IV. List of participanting Centres and Institutions

CENTRIA - Centre for Artificial Intelligence - http://centria.di.fct.unl.pt/

CITI - Centre for Informatics and Information Technologies - http://citi.di.fct.unl.pt/

CEG - Centre for Geological Studies, Universidade Nova de Lisboa – http://www.dct.unl.pt/CEG-novo/CEG.html

CIGA - Centre for Research in Geosciences,Universidade Nova de Lisboa - http://www.ciga.fct.unl.pt/

CQFB - Centre for Chemestry and Biotechnology, Universidade Nova de Lisboa - http://www.cqfb.fct.unl.pt/

CREM - Centre for Microbiology , Universidade Nova de Lisboa - http://www.crem.fct.unl.pt/

CENIMAT - Research Centre in Materials Science, Universidade Nova de Lisboa - http://www.cenimat.fct.unl.pt/

CGE/UE – Centre for Geophysics, Universidade de Évora – http://www.cge.uevora.pt/

IO – Oceanography Institute, University of Lisbon - http://www.io.fc.ul.pt/

GECAD/IPP - Knowledge Engineering and Decision Support Research Center, Porto Polytechnic Institute - http://www.gecad.isep.ipp.pt/Gecad/

ICAM/UE – Mediterranean Agrarian Sciences Institute, Universidade de Évora – http://www.icam.uevora.pt/

DI-FCT/UNL - Department of Computer Science, Universidade Nova de Lisboa - http://www.di.fct.unl.pt/

DI-UE - Department of Computer Science, Universidade de Évora - http://www.di.evora.pt/

V. List of participating Individuals

Note: this list is non-exhaustive, it represents a nucleus which will be extended as appropriate. Alphabetical order was used.

Alexandre Velhinho – Department of Material Sciences FCT/UNL - ajv@fct.unl.pt
Armando Fernandes – CENTRIA – arm.fernandes@gmail.com
Birgit Arnholdt-Schmitt – EU Marie Curie Chair/ICAM UÉvora – eu_chair@uevora.pt
Carlos Damásio – CENTRIA - cd@di.fct.unl.pt
Cecília Gomes – PhD. Student, CITI – mcg@di.fct.unl.pt
Francisco Azevedo – CENTRIA – fa@di.fct.unl.pt
Irene Rodrigues – CENTRIA & UE – ipr@di.uevora.pt
José Cardoso e Cunha – CITI – jcc@di.fct.unl.pt
Lígia Ferreira – CENTRIA & UE – lsf@di.uevora.pt
Luís Moniz Pereira – CENTRIA – lmp@di.fct.unl.pt
Manuel Costa – Department of Environmental Sciences FCT/UNL – manuel.costa@ydreams.com
Nuno Oliveira – PhD. Student, CITI – no@isel.ipl.pt
Paulo Afonso Lopes - PhD. Student, CITI - pal@di.fct.unl.pt
Paulo Quaresma - MSc student, CITI - pjq@di.fct.unl.pt
Pedro Barahona - CENTRIA - pb@di.fct.unl.pt
Pedro Duarte de Medeiros - CITI - pm@di.fct.unl.pt
Pedro Salgueiro – PhD. Student, CENTRIA & UE - pds@di.uevora.pt
Rui Machado - MSc. Student, UE - machador@de.ibm.com
Salvador Pinto Abreu - CENTRIA & UE - spa@di.uevora.pt
Susana Nascimento - CENTRIA - snt@di.fct.unl.pt