

CONTRACT NUMBER 508830

DEISA
**DISTRIBUTED EUROPEAN INFRASTRUCTURE FOR
SUPERCOMPUTING APPLICATIONS**

European Community Sixth Framework Programme
RESEARCH INFRASTRUCTURES
Integrated Infrastructure Initiative

JRA3: Grid-enabling of the Plasma Physics
Code ORB

Deliverable ID: DEISA-D-JRA3-1

Due date : Oct, 31, 2004

Actual delivery date: November 18, 2004

Lead contractor for this deliverable: RZG, Germany

Project start date : May 1st, 2004

Duration: 5 years

Project co-funded by the European Commission within the Sixth Framework Programme (2002-2006)		
Dissemination Level		
PU	Public	
PP	Restricted to other programme participants (including the Commission Services)	X
RE	Restricted to a group specified by the consortium (including the Commission	
CO	Confidential, only for members of the consortium (including the Commission Services)	

Table of Contents

Table of Contents	1
1. Introduction.....	2
1.1 Executive Summary.....	2
1.2 References and Applicable Documents	2
1.3 List of Acronyms and Abbreviations	2
2. General remarks.....	4
3. Grid-enabling of plasma physics simulation code ORB.....	5
3.1 Introduction to ORB code	5
3.2 Code Version and Licensing for DEISA	5
3.3 Scaling of ORB for large number of nodes at RZG.....	6
3.4 ORB code preparation for usage on different DEISA core sites.....	7
3.4.1 Generation of the ORB executable	7
3.4.2 Interactive ORB tests.....	7
3.4.3 ORB preparation for batch usage.....	7
3.5 Portal for ORB code.....	7
3.5.1 UNICORE elements with portal functionality	8
3.5.2 UNICORE configuration for DEISA	9
3.5.3 ORB plug-in.....	10
3.5.4 ORB execution via plug-in	10
3.5.5 Prototype	10
4. European Plasma Physics Community	12

1. Introduction

1.1 Executive Summary

This document “JRA3: Grid-enabling of the Plasma Physics Code ORB” is the 6-month deliverable DEISA-JRA3-1 for Joint Research Activities in Plasma Physics. It describes the several levels of activity necessary to achieve grid-enabling of the ORB code which has been selected as the first candidate. This includes adaptation of the application to the different environments encountered in the DEISA core sites CINECA, IDRIS and RZG, and provision of a portal for ease of use of ORB for new scientists.

1.2 References and Applicable Documents

- [1] <http://www.UNICORE.org/documents/UNICOREPlus-Final-Report.pdf>
- [2] Allfrey, S.J. and Hatzky R.: A revised delta-f algorithm for nonlinear PIC simulation. *Comp. Phys. Commun.*, **154**: 98, 2003
- [3] Hatzky, R., Tran, T.M., Könies, A., Kleiber, R., and Allfrey, S.J.: Energy Conservation in a Nonlinear Gyrokinetic Particle-in-cell Code for Ion-Temperature-Gradient-driven (ITG) Modes in theta-Pinch Geometry. *Phys. of Plasmas*, **9**: 898, 2002
- [4] Villard, L., Allfrey, S.J., Bottino, A., Brunetti, M., Falchetto, G.L., Grandgirard, V., Hatzky, R., Nührenberg, J., Sauter, O., Sorge, S., and Vaclavik, J.: Full Radius Linear and Nonlinear Gyrokinetic Simulations for Tokamaks and Stellarators: Zonal Flows, Applied ExB Flows, Trapped Electrons and Finite Beta. *Nucl. Fusion*, **44**: 172, 2004
- [5] <http://sourceforge.net/projects/unicore/>
- [6] Kim, C.C. and Parker S.E.: Massively Parallel Three-Dimensional Toroidal Gyrokinetic Flux-Tube Turbulence Simulation. *J. Comp. Phys.*, **161**: 589, 2000

1.3 List of Acronyms and Abbreviations

AFS	Andrew File System
AJO	representation of A user JOB, managed and submitted to target system by the NJS
BSS	batch sub-system on the target machine (IBM LoadLeveler)
CA	Certificate Authority; the instance that signs user and server certificates.
CEA	Comité de l'énergie atomique, see http://www-cad.cea.fr
CINECA	Consorzio Interuniversitario, Bologna, http://www.cineca.it
CRPP	Centre de Recherches en Physiques des Plasmas, Lausanne See http://crppwww.epfl.ch
UNICORE configuration	a UNICORE configuration established within DEISA or between some DEISA sites
DEISA UNICORE gateway	a server on a dedicated machine that provides a consistent access to the NJS servers at every DEISA site
DEISA site	partner site of the DEISA consortium

DFN	Deutsches Forschungsnetz, German network provider for the scientific community, http://www.dfn.de
DFN-PCA	DFN Policy Certification Authority, http://www.dfn-pca.de
GUI	Graphical User Interface
JRA	Joint Research Activity
IDB	Incarnation Data Base: the IDB provides site specific descriptions of available computing resources, e.g., such as ORB
IPP	Max-Planck-Institut für Plasmaphysik, Garching, see http://www.ipp.mpg.de
MASS	Mathematical Acceleration Subsystem (MASS): consists of libraries of tuned mathematical intrinsic functions
NJS	Network Job Scheduler: a UNICORE server component that instantiates AJOs or actions (e.g. directory listing on a target machine) by means of a corresponding incarnation data base, schedules jobs and manages job status information and data transfer between the TSI and a gateway or other NJS servers.
RZG	Rechenzentrum Garching, see http://www.rzg.mpg.de
SMP	Symmetric MultiProcessor
TSI	Target System Interface; UNICORE server component which represents the interface to the batch sub-scheduler.
UID	a user's Unique Identifier
UNICORE	UNiform Interface to COmputing Resources, see http://www.unicore.org
UADB	UNICORE User Data Base, maps user certificates to uid
WSMP	Watson Sparse Matrix Package

2. General remarks

In the first year of the DEISA project JRAs in general face the fact that the infrastructure to be built up and provided by the different Service Activities has not yet been fully established. On the other hand, JRAs cannot wait for their complete availability in order to be able to start and proceed with their tasks. This requires extra work to be done by the JRAs in order to prevent severe delays for the achievement of the JRA deliverables which ensure the availability and usability of key applications as soon as the DEISA production environment will be ready. The JRAs have had to implement workaround environments and have provided the SAs with feedback. In particular, JRA3 and JRA1 gave significant support for the installation and DEISA-optimized configuration of a UNICORE test environment at CINECA and RZG.

The objective of this JRA is to grid-enable key applications from plasma physics for usage in DEISA by the European plasma physics community. For this task the ORB code has been selected as the first application. With this code all the different levels of activities have been conducted that are necessary for grid-enabling:

- It must be assured that the code can be executed at the different core sites in an automated way and that it will be able to later use DEISA wide global file systems, so that DEISA-wide job scheduling and rerouting will become possible.
- For new scientists and to generally facilitate code usage and job preparation by new users, a portal should be provided for ease of use.

As a general outlook future activities will include:

- Screening which applications have relevance for the plasma physics community, and preparation of these codes for DEISA wide usage.
- Definition, provision, adaptation and improvements of portals according to the needs of the plasma physics community
- Support for collaborative work of European plasma physicists in the DEISA environment
- Discussions with the plasma physics community about their needs and requests for DEISA

3. Grid-enabling of plasma physics simulation code ORB

3.1 Introduction to ORB code

The ORB code solves gyrokinetic equations pertinent to the study of transport-related instabilities and turbulence. Initiated at CRPP, the ORB code has been substantially upgraded at IPP, and the ongoing code development is made under a close collaborative effort. The ORB code at present exists in several different versions, each specialised to a particular geometry and physics aspect. It covers both stellarator- and tokamak-relevant configurations and solves the problem in the whole plasma domain (full radius). It uses a finite element, "particle-in-cell" (PIC), time evolution approach, and has the unique feature of implementing a statistical optimisation technique that increases the accuracy by orders of magnitude. This has enabled the finding of, for example, long-living zonal flow structures, analogous to those seen in the Jovian atmosphere, and of their crucial importance in determining energy transport.

The numerical quality of the code has been thoroughly assessed. The code performance has been demonstrated to scale well with the number of processing elements (PEs). Less than 5% of the execution time is spent in inter-node communication (with 32 cpus per node) and thus the code has the potential to make the maximum use of parallel-distributed architectures. The ongoing development of the suite of ORB codes aims, in the long term, at simultaneously including the most challenging difficulties: complete 3D geometry, nonlinear, both ion and electron dynamics. At present, even in the simplest geometry (cylinder) and for the simplest gyrokinetic model (adiabatic electrons), accurate nonlinear simulations already require a computing power equal to what is currently available on single platforms. The study of the turbulence behaviour with the increased physical size of the system, as well as all foreseeable and desirable improvements to the physics and/or the geometry, will undoubtedly benefit from, and most certainly be only made possible with a substantially increased computing power.

3.2 Code Version and Licensing for DEISA

Interest for ORB usage has already been expressed by external scientific groups at CEA Cadarache, France. In order to ensure that the ORB simulation code is available to the European plasma physics community in DEISA, it is necessary to clarify the licensing policy with the authors. Negotiations with the developing scientists have been done with the following results:

The DEISA version of ORB consists of the ability to simulate transport-related instabilities and turbulence in theta-pinch geometry including zonal flows; the statistical optimisation technique called "optimised loading" can be used optionally.

The executables of the DEISA version of ORB are free to use in the DEISA project. The only restrictions are that: a) the numerical results, wherever they are published, are indicated as ORB results; b) the following three papers have to be cited: [2], [3] and [4]. For the ease of distribution at different DEISA sites a package was assembled together with a set of test cases. The original structure of the code only included the "optimised loading" procedure as a post processing step. For ease of use for the DEISA users, both parts have been merged to one source code resulting in a single executable.

3.3 Scaling of ORB for large number of nodes at RZG

The ORB code takes into account a hybrid hardware structure consisting of a clustered SMP computer as e.g. a cluster of the IBM 32-way “Regatta” compute node (eServer p690). It uses a concept for parallelisation called domain cloning which has been recently introduced in Ref. [6]. Domain cloning is an extension of one-dimensional domain decomposition and can be implemented very efficiently on a clustered SMP machine. Its key idea is based on the fact that communication inside a node occurs on the shared memory and is faster than communication between nodes which has to pass through the interconnect. Hence, domain decomposition is only done inside a node while copies of the whole physical domain also called clones are distributed over the nodes. Performance test with 16 nodes (512 processors) using the fast interconnect (IBM Federation Switch) have been done at RZG, with accordingly increased problem sizes. Figure 1 shows the speed-up over the number of nodes for both the old results performed with the IBM Colony Switch and the new ones performed with the fast interconnect. The improvement is obvious and an almost perfect scaling up to 16 nodes can be observed.

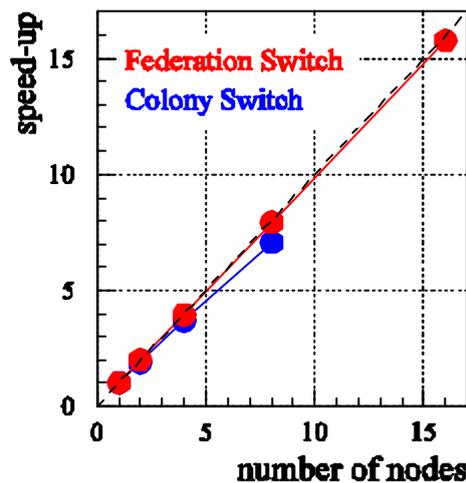


Figure 1 Speed-up of the ORB code as a function of compute nodes. Results for the IBM Colony Switch are plotted in blue, for the fast interconnect (IBM Federation Switch) are plotted in red.

This nearly optimal scaling behaviour on 512 processors (16 nodes) demonstrates that ORB is a serious candidate for even larger runs, for which many more processors will be required. ORB has the potential to run with high efficiency on the largest DEISA sites with parallel batch queues of 1024 processors. Tests at such a large DEISA site (FZJ) are already in preparation.

3.4 ORB code preparation for usage on different DEISA core sites

Application specific porting work had to be done for the adaptation to the different environments .

3.4.1 Generation of the ORB executable

ORB performance significantly benefits from linking to the optimized numerical libraries WSMP and MASS which are both available at RZG. MASS is also available at IDRIS and CINECA, WSMP not. Furthermore, environment variables are not consistently defined across different sites. Therefore site-specific makefiles had to be generated which will be updated as soon as environments will be fully unified with the module concept of SA4. A specific ORB module will be provided by JRA3 based on the site specific modules provided by SA4.

3.4.2 Interactive ORB tests

Comparative tests for interactive parallel execution have been done at RZG, CINECA and IDRIS with the (IBM specific) poe command. Since both native and site-specific modified IBM poe environments were encountered, poe commandline options had to be adapted for different sites.

3.4.3 ORB preparation for batch usage

Batch usage will be the standard operation mode of DEISA. The goal will be to have a single batch script that works at different DEISA sites. This is a prerequisite for future scheduling and in particular rerouting of jobs across DEISA.

However, system configurations of the batch scheduler (IBM LoadLeveler) differ from site to site due to different underlying concepts (e.g. "class" concept vs. "requirements" concept). As a workaround site-specific Loadleveler batch scripts have been generated. Unification of the differing concepts would be the recommended solution. If that cannot be achieved a tool for automatic batch script adaptation has to be provided.

3.5 Portal for ORB code

New users and scientists not well familiar with the instrumentation of the ORB code can benefit from a portal which provides an uniform access to the DEISA infrastructure. Such a portal has to facilitate the application handling, i e. job preparation, job submission and output handling. This is a way to hide details of the infrastructure from the user.

Such functionality is principally provided by special features of the UNICORE software package, so-called plug-ins for the UNICORE client. In order to be able to test the suitability of such features for ORB, a complete UNICORE test environment had to be built up at two DEISA sites at least. This was done in a joint effort of SA3, JRA3 and JRA1 for CINECA and RZG. SA3 focuses more on middleware functionality aspects in general. JRA3 (and JRA1) were especially interested in the plug-ins for the UNICORE client. This plug-in technology can be further developed in the scope of the JRAs.

3.5.1 UNICORE elements with portal functionality

UNICORE [1] provides a seamless interface for preparing and submitting jobs to various computing resources. UNICORE is designed according to a three-tier-model, represented by:

- ? a UNICORE client that facilitates the preparation and administration of jobs (so-called AJOs) during their workflow stages by means of a GUI;
- ? a Network Job Scheduler (NJS), i.e. a server which provides site specific information on the computing resources and available applications. The NJS manages the AJOs for the submission, collects and keeps their results after execution on the target machine;
- ? a Target System Interface (TSI) that represents the interface to the batch sub-scheduler (BSS) on a target machine.

Among other grid relevant functionalities, the modular plug-in technology of the client makes UNICORE a serious candidate for providing the necessary portal functionality. Therefore, UNICORE has been considered to test and demonstrate the grid-enabling of the application ORB.

In order to be able to investigate and evaluate the potential value of the UNICORE client for portal suitability for ORB, a UNICORE test environment had to be installed. Corresponding work had to be done on each of the three UNICORE layers mentioned above:

- ? a client plug-in has been developed
- ? the incarnation data bases (IDB) of the NJS has been configured appropriately
- ? the environments on the target machines (including site specific adaptations of the TSI) have been organised

3.5.2 UNICORE configuration for DEISA

The SA3 team focuses on the general middleware aspects of UNICORE. JRA3 (and JRA1) were particularly interested in the so-called plugins for the UNICORE client. In close cooperation with Paolo Malfetti and Andrea Vanni from CINECA, a UNICORE configuration as depicted by figure 1 has been established. This configuration allows the submission of ORB jobs to either RZG or CINECA HPC resources for investigations. As a first step the test installation has been done including RZG and CINECA with the option to extend this scalable infrastructure and integrate other DEISA sites (e.g. IDRIS and FZJ).

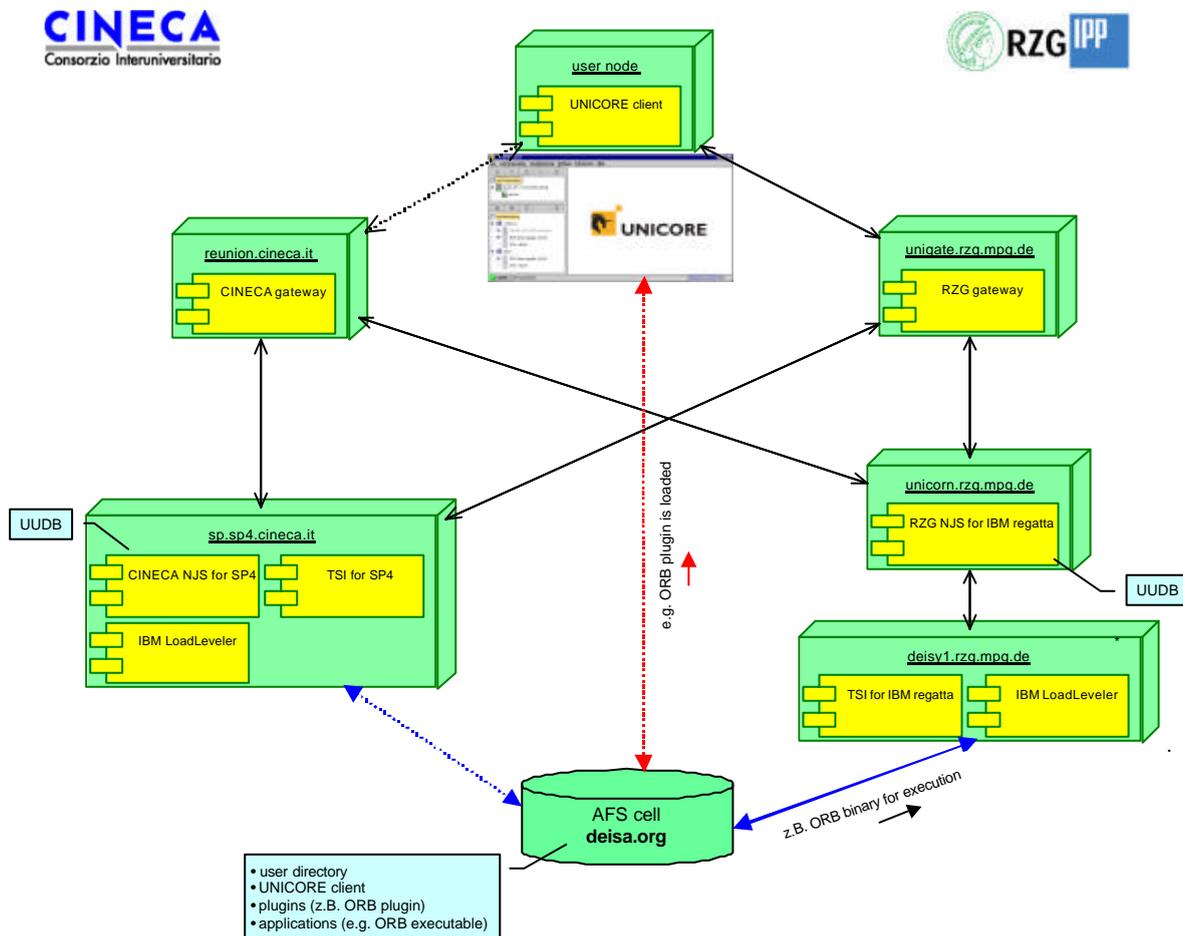


Figure 2 UNICORE configuration for cross-site tests with ORB. The UNICORE client, the ORB plug-in and the ORB executable can be distributed via AFS servers which are already accessible by RZG, CINECA and IDRIS. A UNICORE client which is running on a node with AFS access can load the ORB plug-in directly from there. Dotted connections are optional.

UNICORE client 5.1.2, NJS server 4.0.3, gateway server version 4.0.1 and TSI server 4.0.4 from <http://unicore.sourceforge.net> [5] have been deployed. The configuration has been tested successfully by submitting ORB jobs, prepared with the ORB plug-in, between CINECA and RZG.

In order to be able to perform these tests, however, principal problems with certificates for UNICORE had to be solved. The general DEISA certificate policy is still in the process of being developed, from the European level down to national and further down to site level (with SA5 being in charge).

However, in order to meet the JRA3 requirements during the start phase of the DEISA project, these early tests with ORB in the framework of the UNICORE infrastructure (see figure 1) had to be performed with preliminary self-signed certificates that are temporarily accepted between CINECA and RZG.

In order to ensure that the user and server certificates will be accepted also in a further enlarged DEISA infrastructure, RZG managed to receive DFN-PCA compliant certificates. In parallel a process has been initiated together with FZJ and DFN that aims at the membership of DFN in the EU Grid PMA (Policy Management Agreement).

3.5.3 ORB plug-in

The ORB application is either located in a separate directory on the target machine at each site or in a global directory. It is invoked by the TSI and receives appropriate input data controlled by the UNICORE client via the NJS.

The advanced plug-in for ORB has been developed regarding to ease of use.

An ORB specific form sheet has been developed (see fig. 2). It already contains reasonable default values which are user specific and can be individually modified and stored. The form can now also be backfilled with the content of pre-existing input files, since a parser for the input file has been written.

This setup is assembled to an ORB compliant parameter file that is transferred to the TSI before the ORB executable is submitted to the BSS on the target machine.

3.5.4 ORB execution via plug-in

The ORB plug-in was investigated and analysed for principal suitability for usage with ORB and potentially other plasma physics applications. It has been tested on both server installations at CINECA and RZG, employing UNICORE clients under Linux and Windows.

Authentication by certificates, job preparation, submission, execution and result retrieval have been tested. The functional components are verified by a number of test cases.

3.5.5 Prototype

The ORB plug-in for UNICORE client, in combination with the established UNICORE test infrastructure, represents a prototype environment in which ORB can now be used by other DEISA users. Production runs, however, will require further achievements from the other DEISA Service Activities, and, of course, the transfer of the functionality of the test infrastructure to the production systems, a goal only after 12 months of project start.

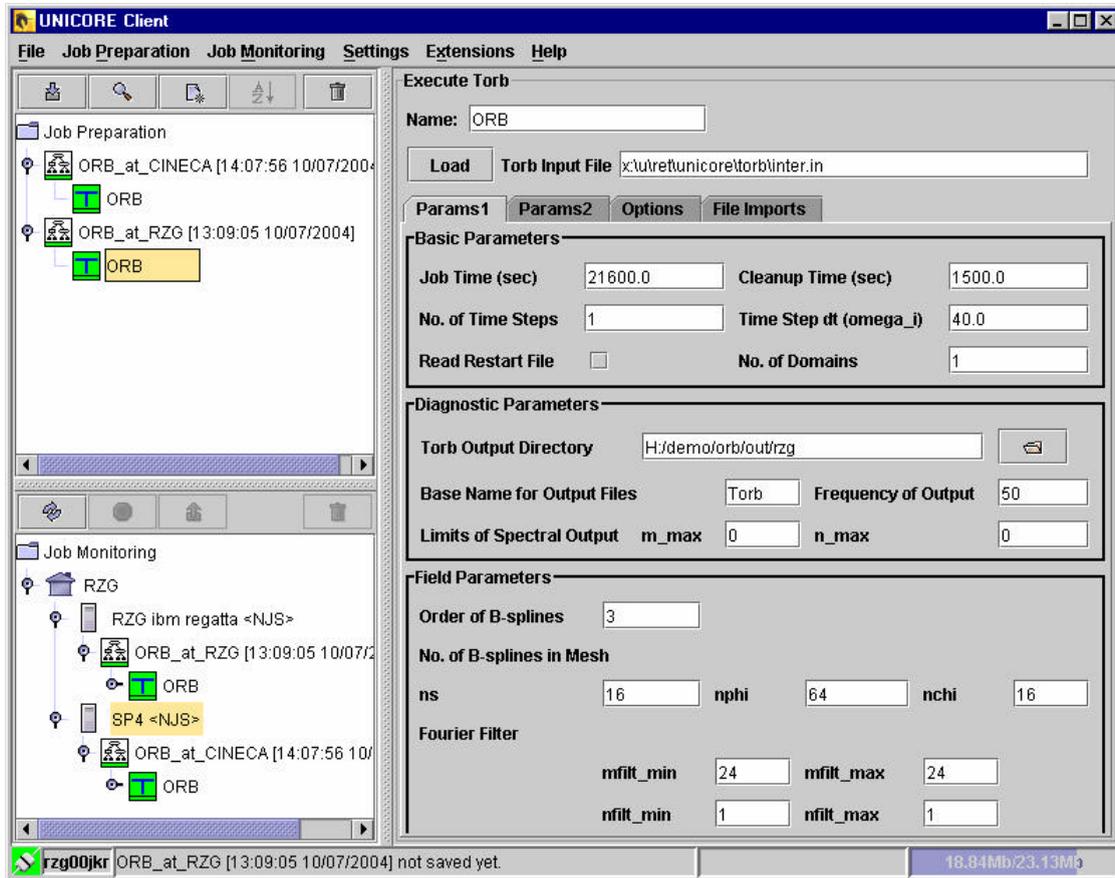


Figure 3: example of the ORB plug-in and a demonstration of the job submission to the two DEISA sites CINECA and RZG for test purposes.

Window upper left: Job preparation panel
 Window lower left: Panel for job monitoring and status control
 (submitted, pending, running, finished, aborted)
 Window right: Panel for parameter details when upper left window is active

4. European Plasma Physics Community

This DEISA JRA in plasma physics has been announced to the European scientific community at the *Theory of Fusion Plasmas Joint Varenna-Lausanne International Workshop*, Aug/Sep 2004 (<http://varenna-lausanne.epfl.ch>).

As a first feedback interest has been announced by two scientific groups at CEA Cadarache to perform benchmarks with the ORB code version of JRA3. One of these groups has been in close collaboration with CRPP Lausanne to develop a Semi-Lagrangian code to simulate transport-related instabilities and turbulence.

The goal is to benchmark equivalent physical simulations with quite different numerical methods (the Semi-Lagrangian code is discretising the physical phase space with a mesh while the ORB code is based on a Monte Carlo approach). Such benchmarks are very important as they give the chance to cross-check the results with an independent numerical method. In particular, systematic errors due to the applied numerical method can be identified in this way. This is of special interest because turbulence simulations are already so complicated that analytical benchmarks are usually not available and can be only replaced partly by convergence studies of numerical parameters such as mesh size and particle number.