

CONTRACT NUMBER 508830

**DEISA**  
**DISTRIBUTED EUROPEAN INFRASTRUCTURE FOR  
SUPERCOMPUTING APPLICATIONS**

**European Community Sixth Framework Programme**  
RESEARCH INFRASTRUCTURES  
Integrated Infrastructure Initiative

JRA3: Provision of a portal for simulation code suite (ORB),  
preparation for multi-site usage

Deliverable ID: DEISA-D-JRA3-2

**Due date: April, 30, 2005**  
**Actual delivery date: May 15, 2005**  
**Lead contractor for this deliverable: RZG, Germany**

**Project start date: May 1<sup>st</sup>, 2004**  
**Duration: 5 years**

Project co-funded by the European Commission within the Sixth Framework Programme (2002-2006)		
Dissemination Level		
<b>PU</b>	Public	
<b>PP</b>	Restricted to other programme participants (including the Commission Services)	<b>X</b>
<b>RE</b>	Restricted to a group specified by the consortium (including the Commission	
<b>CO</b>	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. Provision of a Portal for the ORB code suite.....	4
3. TORB code preparation for multi-site usage .....	7
4. Scaling of TORB for extreme computing.....	10
5. European Plasma Physics Community .....	14

## 1. Introduction

### 1.1 Executive Summary

This document “JRA3: Provision of a portal for simulation code suite (ORB), preparation for multi-site usage” is the 12-month deliverable DEISA-JRA3-2 for Joint Research Activities in Plasma Physics. It describes the several levels of activity necessary to achieve the described goal, including portal functionality, multi-site usage, preparation of scalability for extreme computing, and further activities to expand the scientific user community.

### 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

BSS	Batch Sub-System on the target machine
CEA	Comité de l'énergie atomique, see <a href="http://www-cad.cea.fr">http://www-cad.cea.fr</a>
CINECA	Consorzio Interuniversitario, Bologna, <a href="http://www.cineca.it">http://www.cineca.it</a>
CRPP	Centre de Recherches en Physiques des Plasmas, Lausanne See <a href="http://crppwww.epfl.ch">http://crppwww.epfl.ch</a>
DEISA site	partner site of the DEISA consortium
JRA	Joint Research Activity
GPFS	Global Parallel File System
IPP	Max-Planck-Institut für Plasmaphysik, Garching, see <a href="http://www.ipp.mpg.de">http://www.ipp.mpg.de</a>
ECMWF	European Centre for Medium-Range Weather Forecast, see <a href="http://www.ecmwf.int">http://www.ecmwf.int</a>
EPCC	Edinburgh Parallel Computing Centre, see <a href="http://www.epcc.ed.ac.uk">http://www.epcc.ed.ac.uk</a>

RZG	Rechenzentrum Garching, see <a href="http://www.rzg.mpg.de">http://www.rzg.mpg.de</a>
SMP	Symmetric MultiProcessor
TSI	Target System Interface: UNICORE server component which represents the interface to the batch sub-scheduler
UNICORE	UNiform Interface to COmputing Resources, see <a href="http://www.unicore.org">http://www.unicore.org</a>

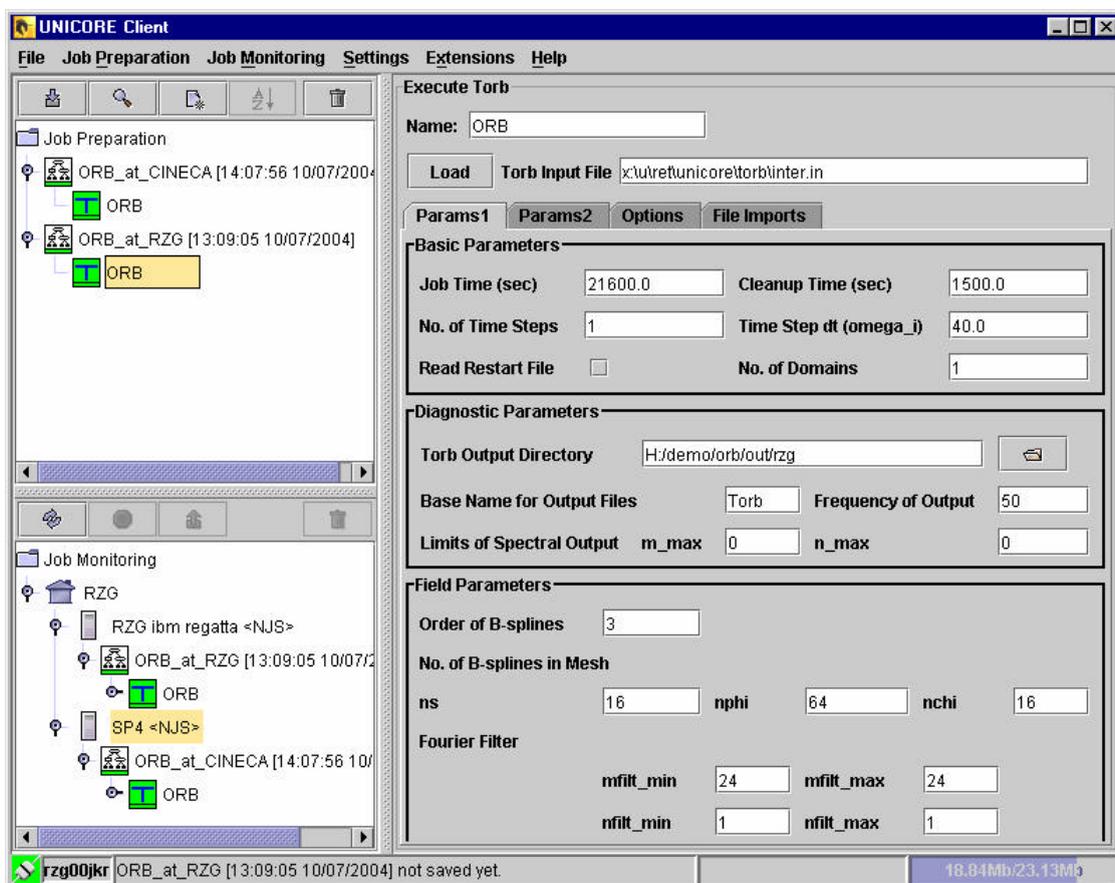
## 2. Provision of a Portal for the ORB code suite

ORB code suite is a set of related codes, solving identical gyrokinetic equations for different geometrical configurations. For the JRA3 activities with ORB as a prototype, the Theta-Pinch geometry had been selected for DEISA, as already explained in the JRA3 6 months deliverable DEISA\_D-JRA3-1. This version will from now on be referred as TORB, since new versions are currently under development.

New users and scientists not familiar with the instrumentation of the ORB code suite should benefit from a portal which provides a 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 can principally be provided by special features of a part of the UNICORE software package, so-called plug-ins for the UNICORE client. This plug-in technology can be further developed in the scope of the JRAs.

As a first step, an advanced plug-in for TORB has been developed for the UNICORE client, focusing on ease of use. A TORB-specific form sheet has been developed. 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.



**Fig. 1** TORB plug-in for the UNICORE client

This setup is assembled to a TORB-compliant parameter file that is transferred to the relevant Target System Interface (TSI) before the TORB executable is submitted to the Batch Sub-System (BSS) on the target machine.

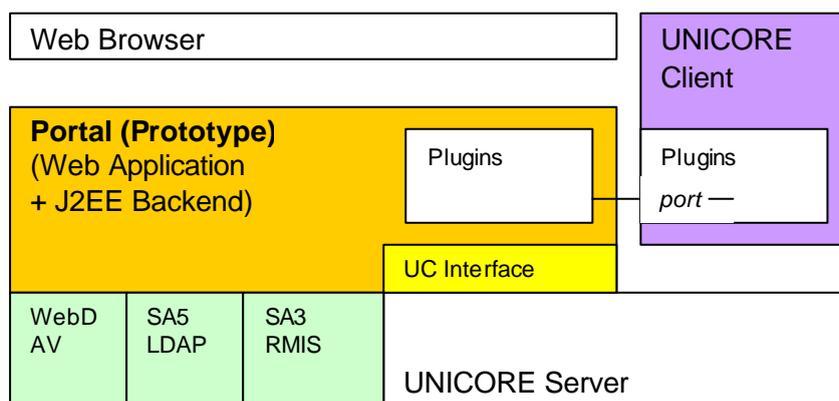
Authentication by certificates, job preparation, submission, execution and result retrieval have been tested. The functional components are verified by a number of test cases.

This first basic approach for the portal functionality has already been described in the 6-month deliverable of JRA3 (DEISA\_D-JRA3-1.doc), where the prototype has also been presented.

Two subsequent activities have been carried out since then. First, the existing prototype was converted to a production version that can readily be used on DEISA core sites, together with the UNICORE client and the associated ORB plug-in.

Secondly, a more general approach (than Unicore client plus plug-in) to portal functionality is on the way, necessary also for applications in JRA1 for materials science. This new approach and the reasons for it are described in detail in the 12 month deliverable of JRA1 (DEISA\_D-JRA1-2.doc).

Essentially, the DEISA infrastructure access will be transferred from the UNICORE client to a new portal (web application) at the server side, directly interfaced to the UNICORE server. This portal can be directly accessed with a standard web browser: no further client software installation will be needed at user side. The following diagram, taken from DEISA\_D-JRA1-2.doc, illustrates the situation.

**Fig. 2**

DEISA portal options for JRA3 (and JRA1)

Right side: Unicore client plus plug-in (basic, UNICORE client compatible solution)

Left side: New, more general approach with web application interface to UNICORE server, including plug-in(s)

The new general portal approach is already under development as a prototype with TORB, the TORB plug-in has been adapted and TORB could, in principle, be started in DEISA core sites using a simple standard web browser.



**Fig. 3**

Design of the TORB plug-in, a new web application/portal, accessible via the web browser Mozilla.

The plug-in for the second plasma physics code, GENE, which is already under development, will be provided soon in the UNICORE client compatible version, and later on for the new portal.

### 3. TORB code preparation for multi-site usage

The term multi-site usage could be used in different contexts:

- ? synchronous multi-site usage of coupled applications;
- ? a single application being spread over more than one site;
- ? asynchronous or sequential multi-site usage in a workflow of different functions;
- ? a continuous series of jobs and optional post-processing and/or visualization of results.

Since synchronous single application distribution over different sites ("metacomputing") has not become a major operational model of DEISA, but job re-routing across sites will be a focus, we have addressed and tested sequential multi-site usage. The aspect of being able to use large resources for a single application (one original motivation for single application metacomputing in the late nineties) is addressed in chapter 4. Scalability expansion has become a new perspective with the availability of more than a thousand processors in a single site. Access to such large-scale computing via DEISA is already realized with ECMWF becoming a full partner, and will soon also be available at BSC and LRZ next year.

Here we have focused on starting TORB execution at one site, writing a restart file, doing a continuation run at the same site, continuing the simulation at a second site, and starting a post-processing application for visualization at a third site.

Such a usage type has been prepared and conducted involving compute systems at three of the DEISA core sites, IDRIS, FZJ and RZG. CINECA systems will be included in the next test series.

TORB multi-site usage could already benefit from the new global file system multi-cluster GPFS among the DEISA core sites, and has also been used as the first real application from a JRA to test the setup.

**1. Start of TORB simulation run****2. Continuation of TORB sim. run****3. Start of TORB post-processing run****Figure 4**

Multi-site usage of TORB involving the new global file system GPFS among DEISA core sites: start of a TORB simulation at IDRIS, continuation of simulation at FZJ, start of post-processing application at RZG.

**Multi-site tests with TORB**

TORB has been tested on the dedicated systems where early versions of the multi-cluster GPFS were available. Typically, eight processors have been used for a problem size of 12 GB, a simulation with 20.000.000, particles, and reading/writing restart files of size 1,875 Mb.

The first TORB simulation was started at IDRIS (Orsay, France) writing/reading data to/from GPFS disks located at IDRIS (three continuation runs were done).

The following lines were taken from the log file, referring to restart file I/O:

```
Particle out wall clock time = 8.3466280E+00 => 224.6MB/s
```

```
Particle in wall clock time = 9.6171534E+00 => 195.0MB/s
```

```
Particle out wall clock time = 8.4093769E+00 => 223.0MB/s
```

```
Particle in wall clock time = 9.9078562E+00 => 189.2MB/s
```

```
Particle out wall clock time = 8.5982108E+00 => 218.1MB/s
```

The TORB simulation was then continued at FZJ (Juelich, Germany), writing/reading data to/from GPFS disks located at IDRIS (again three continuation runs were done).

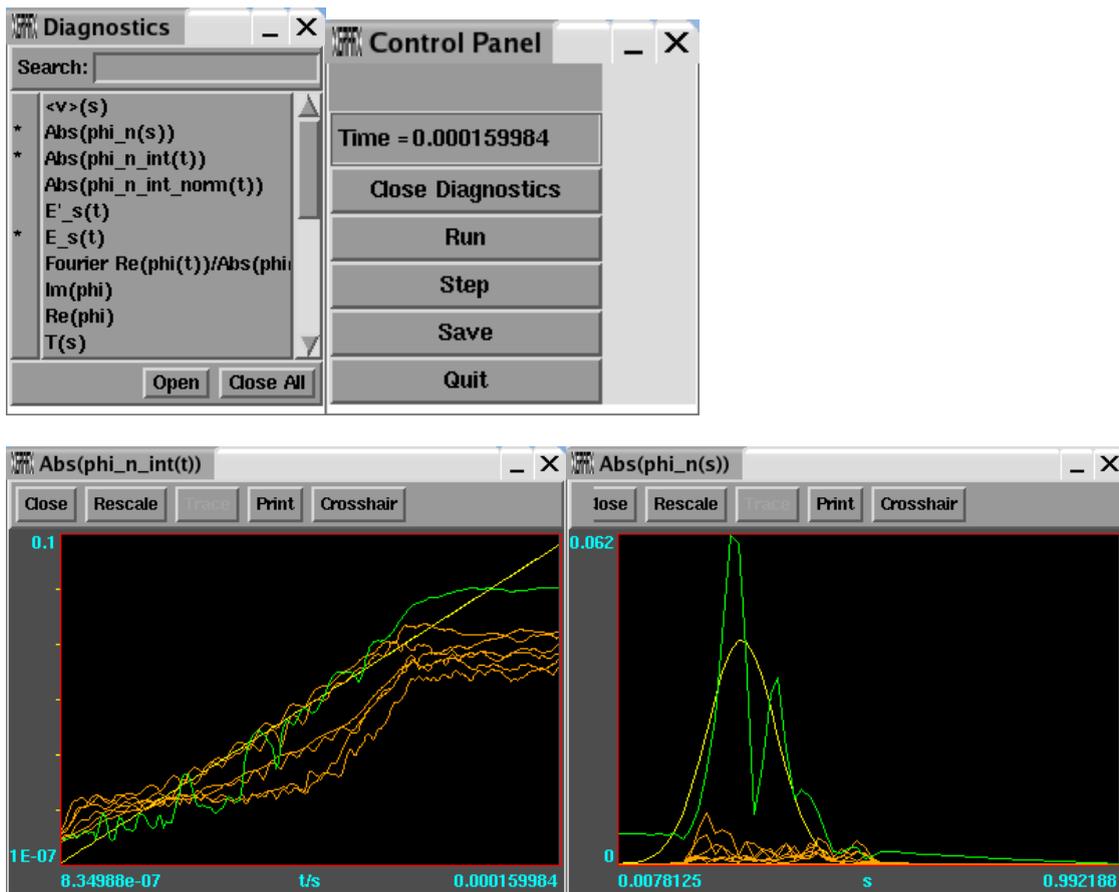
The following lines were taken from the log file, referring to restart file I/O:

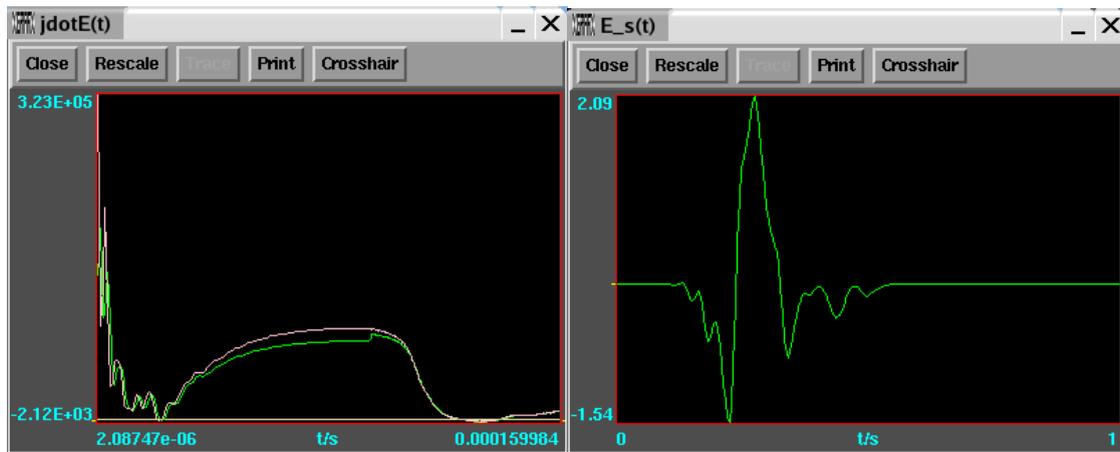
```
Particle in wall clock time = 1.9522469E+01 => 96.0MB/s
Particle out wall clock time = 3.6178846E+01 => 51.8MB/s
```

```
Particle in wall clock time = 1.7238927E+01 => 108.8MB/s
Particle out wall clock time = 4.2755378E+01 => 43.9MB/s
```

```
Particle in wall clock time = 1.6924593E+01 => 110.9MB/s
Particle out wall clock time = 3.5461993E+01 => 52.9MB/s
```

In a third step, the post-processing tool ANGY for TORB output data evaluation was started at RZG, accessing GPFS data located at IDRIS. The six screenshots in Fig. 5 document this post-processing step.





**Fig. 5**

Visualization of various TORB simulation output parameters with analysis tool ANGY. Screenshots of 6 different windows are displayed.

For its graphical user interface, the data analysis tool ANGY uses the graphical software package xgrafx written by the Plasma Theory and Simulation Group (PTSG) at UC Berkeley (URL: <http://langmuir.nuc.berkeley.edu/pub/codes/xgrafx/>). The flow chart of ANGY was adapted to upgrade the xgrafx package from version 1.94 to the recent version 2.60.

#### 4. Scaling of TORB for extreme computing

The nonlinear particle-in-cell code TORB is one of the potential candidates for the DEISA Extreme Computing Initiative. It uses a Monte Carlo particle approach to simulate the time evolution of turbulent field structures in fusion plasmas. In such simulations, a very large number of Monte Carlo particles is needed to simulate large physical domains with a very low level of statistical noise for long simulation intervals. The large amount of memory used for storing these hundreds of millions of particles is only made available by massively parallel computers with distributed memory architectures.

Although Monte Carlo codes have in general good scalability properties on parallel computers, it was not clear how TORB would scale up to thousands of processors. Due to Amdahl's law it is a nontrivial task to run codes efficiently on thousands of processors: such extreme scalability can be only reached by a small number of codes.

As a first step, the scaling properties of the TORB code were documented up to 512 processors which was limited by the largest available batch queue of the Garching computing centre RZG (see deliverable DEISA\_D-JRA3-1.doc). As a next step measurements on 1024 processors could be done at the supercomputer of the HPCx Consortium of which EPCC is a member. But finally it had been possible to use up to 2048 processors of the supercomputer of the European Centre for Medium-Range Weather Forecast (ECMWF.). Performing such benchmarks on thousands of processors belonging to a clustered symmetric-multiprocessor (SMP) computer was made possible since all named computer centres are members of the DEISA project.

Besides the usual changes to the batch scripts, some minor adaptations of the output format were made to handle processor numbers larger than 999 processors. More significantly, a major change became necessary to handle a number of Monte Carlo particles larger than  $2^{31}=2,147,483,648$ . For the first time the total memory of approximately 2 TByte available gave the opportunity to use 3.2 billion particles within a 2048 processor run of the TORB code which exceeded by far the  $2^{31}$  limit. The  $2^{31}$  particle limit is critical for the programming structure of the TORB code as the standard 4-Byte Integer representation of FORTRAN can not handle larger integer numbers without causing an overflow. Hence the parts of the code relying on 8-Byte Integer arithmetic had to be adapted. A general change to 8-Byte integers would have been more laborious because of the involved changes of the Message Passing Interface (MPI) library interfaces programmed originally in TORB for 4-Byte Integers.

The time spent in sequential regions is reduced to a minimum for the TORB code. However, due to the basic structure of the particle-in-cell simulation each Monte Carlo particles does not act completely independently from the others. Instead, the particles interact with the field resulting from a collective charge-assignment process so a parallel communication is unavoidable. An efficient implementation of this parallel communication is only possible on the deep algorithmic level.

In 2000, a new approach to parallelization, domain cloning, was presented by C.C. Kim, S.E. Parker, *Journal of Computational Physics* 161 (2000) 589-604.

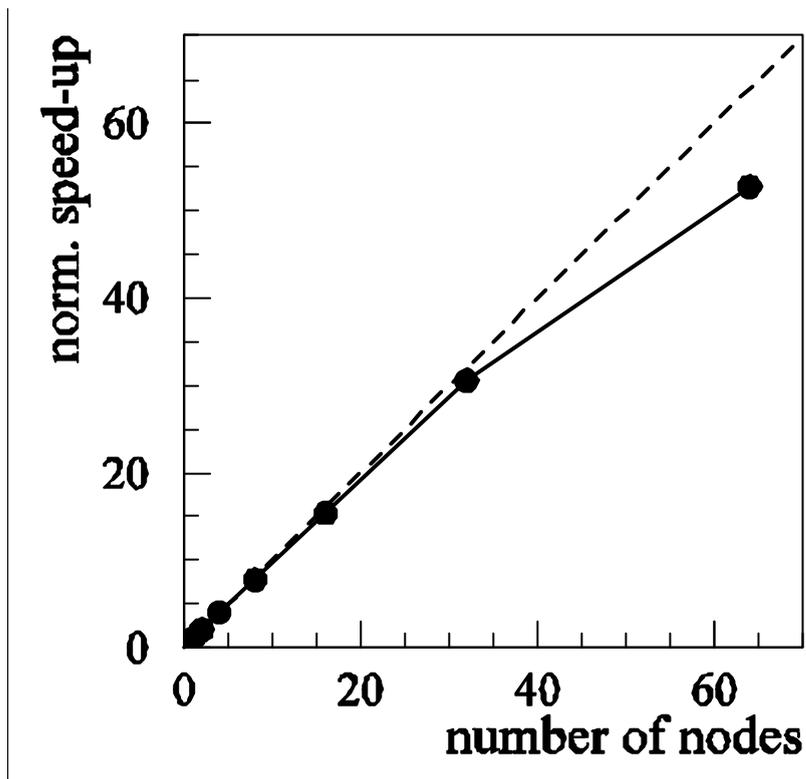
Domain cloning is an additional layer of parallelization, a supplement to one-dimensional domain decomposition, which gives the opportunity to optimize the scaling property of particle-in-cell codes such as the TORB code. In the past this parallelization concept was implemented into TORB at the Garching computing centre. Recently, in collaboration with the DEISA initiative, a further adaptation focused on clustered symmetric-multiprocessor computers, addressing the question of whether it would be possible to use the domain cloning concept up to thousands of processors. As a result, a paper entitled "Domain Cloning for a Particle-in-Cell (PIC) Code on a Clustered Symmetric-Multiprocessor (SMP) Computer" was submitted to the journal *Parallel Computing*.

As mentioned above, the final results have been achieved on one of the two IBM Cluster 1600 supercomputer systems at the High Performance Computing Facility of ECMWF. Each separate cluster comprises 70 pSeries 690+ servers (compute nodes), each of which has 32 1.9 GHz Power 4+ CPUs. Up to 64 compute nodes with 32 GB memory each have been available for the largest simulation. The first 1024 processors runs were conducted on the HPCx system comprising 50 IBM pSeries 690+ nodes, i.e. 1600 Power 4+ processors at 1.7 GHz, with a total of 1.6 TBytes of memory.

Starting with 50 million particles for a single node simulation the number of particles increases linearly with the number of nodes. The maximum number of nodes used is 64 which corresponds to a total number of 3.2 billion Monte Carlo particles in the simulation. Figure 6 shows the speed-up with normalized simulation size as a function of the number of nodes  $n$ . The scaling proves to be excellent for up to 32 nodes. Only a further doubling of the number of nodes to  $n=64$  gives a significant decrease of the speed-up to  $53/64$ . Thus, the intensive allreduce communication (global sum) over the clones finally diminishes the very good scaling properties of the domain cloning concept. Nevertheless, a total flop rate of 1.3 Tflop/s was reached on 64 nodes (2048 PEs) surpassing the Teraflop performance threshold.

After a careful study of the parallelization concept of the TORB code it has been possible to adapt the algorithmic structure in such a way that TORB can use efficiently the powerful hardware made available by the DEISA initiative.

Hence, because of its scientific excellence and now also its extraordinary parallel scaling properties, TORB is a serious candidate for the DEISA Extreme Computing Initiative.



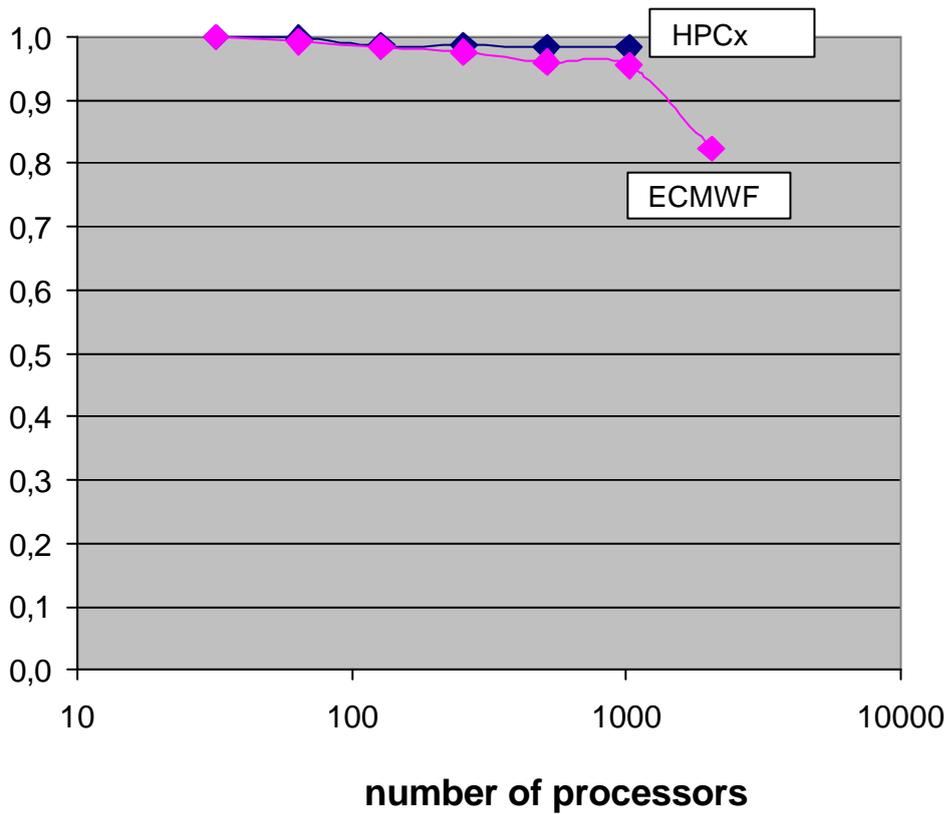
**Figure 6**

The speed-up of the ORB code with normalized simulation size vs. the number of compute nodes (with 32 processors each) involved. The problem size is scaled (increased) linearly with the number of nodes. The dashed line depicts ideal scaling. With 2048 processors (64 nodes), a parallel efficiency of 82 % was achieved, resulting in a performance of 1.3 TFlop/s.

# procs	Parallel efficiency at EPCC	Parallel efficiency at ECMWF
<b>32</b>	1,000	1,000
<b>64</b>	0,998	0,993
<b>128</b>	0,986	0,984
<b>256</b>	0,986	0,975
<b>512</b>	0,983	0,960
<b>1024</b>	0,983	0,955
<b>2048</b>		0,824

**Table 1**  
TORB scaling measurements at EPCC (HPCx system) and ECMWF

**Parallel efficiency**



**Fig 5**  
Parallel efficiency of TORB scaling:  
Measurements at EPCC (HPCx system) and ECMWF

## **5. European Plasma Physics Community**

Preparation of the ORB code suite for adequate usage in DEISA helps with the expansion of the European user community. Developers are located at: IPP, Greifswald, Germany; IPP, Garching, Germany; and CRPP, Lausanne, Switzerland. With the help of principal investigators of this JRA from Germany and Switzerland, contacts to two interested groups at Cadarache in France have been initiated. Now additional interested groups from Spain have been identified (CIEMAT, Madrid) who could benefit from this DEISA JRA effort in trans-national European collaborations. This has also to be considered in the context of BSC, Barcelona, becoming a DEISA partner, and plans have already been agreed to port TORB to that new machine, currently listed as Europe's most powerful supercomputer.