

DEISA overview: project status, strategies and perspectives

Towards cooperative extreme computing in Europe

Victor Alessandrini
IDRIS - CNRS
va@idris.fr

Achim Streit
FZJ
a.streit@fz-juelich.de



Table of contents

- **1 - Project Objectives**
- **2 - Infrastructure overview**
- **3 - Global strategies**
- **4 - Infrastructure organization**
- **5 - Running the infrastructure: services**
- **6 - Running the infrastructure: applications**
- **7 - Prospective issues**
- **8 - Conclusions**

1 - DEISA objectives

- *To enable Europe's terascale science by the integration of Europe's most powerful supercomputing systems.*
- *Enabling scientific discovery across a broad spectrum of science and technology is the only criterion for success*
- **DEISA is an European Supercomputing Service built on top of existing national services. This service is based on the deployment and operation of a persistent, production quality, distributed supercomputing environment with continental scope.**
- **The integration of national facilities and services, together with innovative operational models, is expected to add substantial value to existing infrastructures.**
- **Main focus is High Performance Computing (HPC).**

2 - Infrastructure overview

- Description of the hardware infrastructure

DEISA Sites



BSC	<i>Barcelona Supercomputing Centre</i>	Spain
CINECA	<i>Consortio Interuniversitario per il Calcolo Automatico</i>	Italy
CSC	<i>Finnish Information Technology Centre for Science</i>	Finland
EPCCC/HPCx	<i>University of Edinburgh and CCLRC</i>	UK
ECMWF	<i>European Centre for Medium-Range Weather Forecast</i>	UK (int)
FZJ	<i>Research Centre Juelich</i>	Germany
HLRS	<i>High Performance Computing Centre Stuttgart</i>	Germany
IDRIS	<i>Institut du Développement et des Ressources en Informatique Scientifique - CNRS</i>	France
LRZ	<i>Leibniz Rechenzentrum Munich</i>	Germany
RZG	<i>Rechenzentrum Garching of the Max Planck Society</i>	Germany
SARA	<i>Dutch National High Performance Computing and Networking centre</i>	The Netherlands

THE DEISA SUPERCOMPUTING GRID



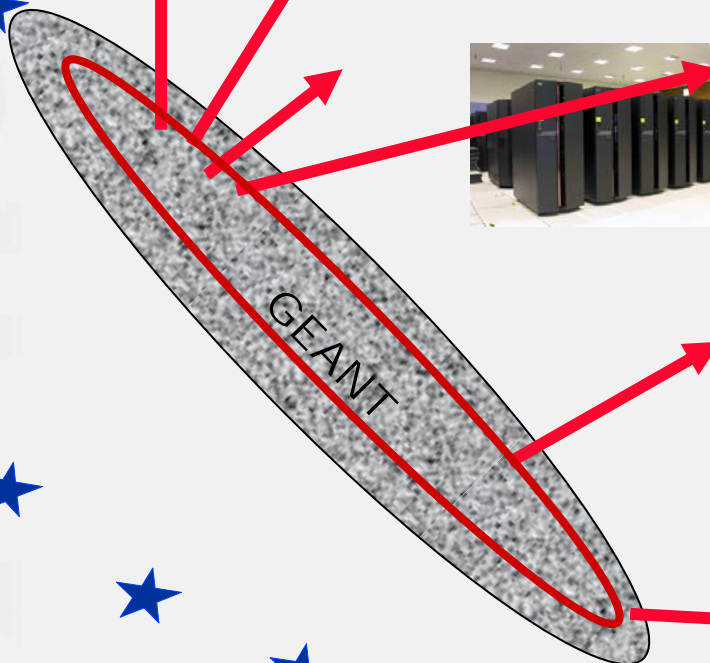
**AIX distributed
super-cluster**



**Vector systems
(NEC, ...)**



**Linux systems
(SGI, IBM, ...)**



The DEISA supercomputing environment

(21.900 processors and 145 Tf in 2006, more than 190 Tf in 2007)

- IBM AIX Super-cluster
 - FZJ-Jülich, 1312 processors, **8.9 teraflops peak**
 - RZG – Garching, 748 processors, **3.8 teraflops peak**
 - IDRIS, 1024 processors, **6.7 teraflops peak**
 - CINECA, 512 processors, **2.6 teraflops peak**
 - CSC, 512 processors, **2.6 teraflops peak**
 - ECMWF, 2 systems of 2276 processors each, **33 teraflops peak**
 - HPCx, 1536 processors, **11 teraflops peak**
- BSC, IBM PowerPC Linux system (MareNostrum), 4864 processeurs, **40 teraflops peak**
- SARA, SGI ALTIX Linux system, 416 processors, **2.2 teraflops peak**
- LRZ, SGI ALTIX Linux system, 4096 processeurs, **26 teraflops peak**
- HLRS, NEC SX8 vector system, 576 processors, **12.7 teraflops peak**
- **Systems interconnected with dedicated 1Gb/s network – currently upgrading to 10 Gb/s – provided by GEANT and NRENs**

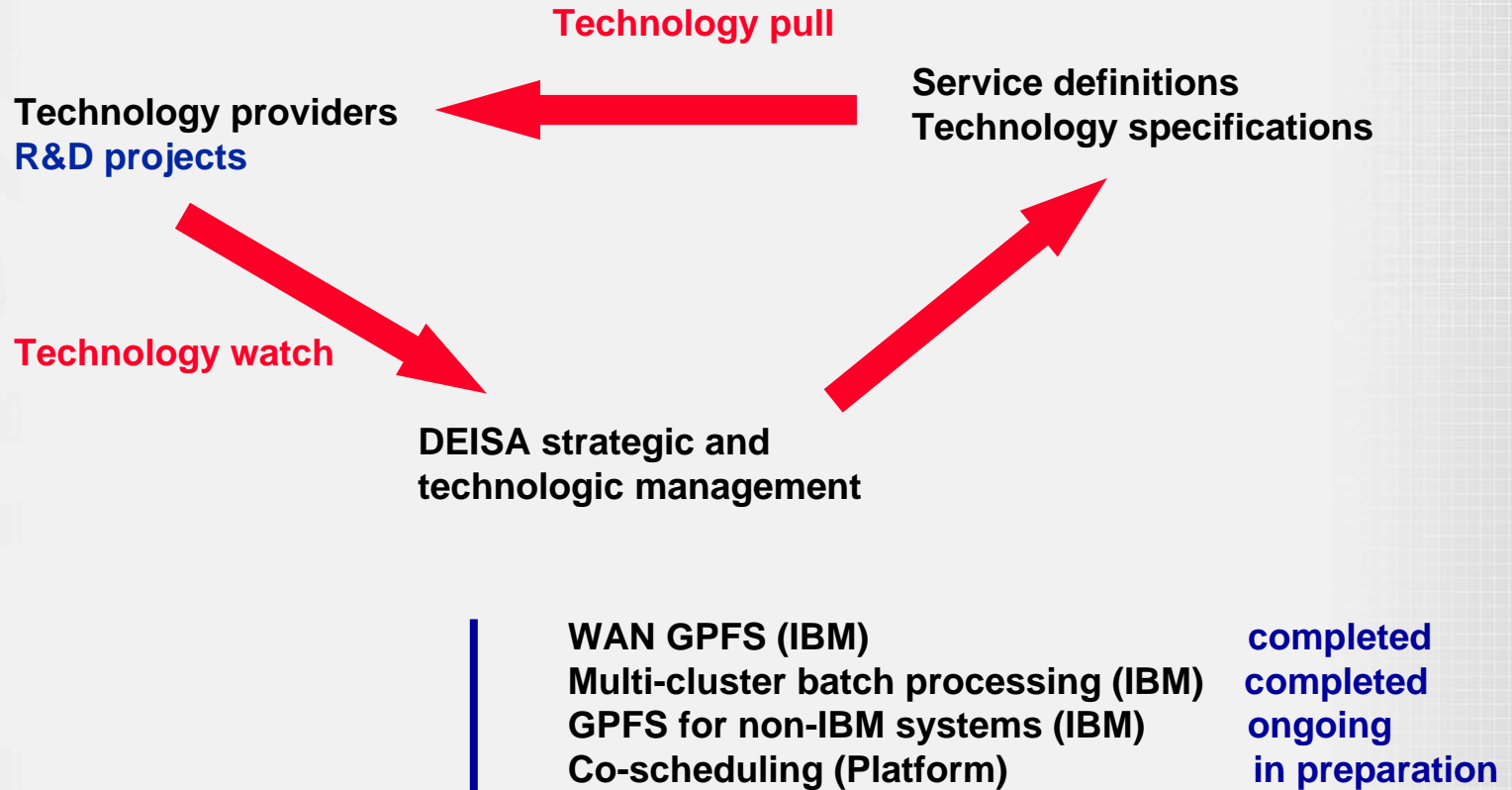
3 - Global strategies

- **General ideas about the DEISA computing strategies and its positioning in the European computational ecosystem.**

3a - Strategic vision

- DEISA is based on a **deep integration and tightly coupled operation** of high end computational resources, rather than a loose federation model.
- Scientific impact (enabling new science) is the only criterion for success.
- Integration of IT systems is mainly a strategic issue. Technology choices follow from the business and operational models of virtual organizations.
- This is why DEISA puts forward an innovative vision for the integration of high end computing systems: promoting cluster software (global file systems, batch managers) to distributed super-cluster middleware (distributed global file systems, multi-cluster batch managers).
- DEISA technology choices are fully open. DEISA is not tied to any specific pre-established technology.

3b - The technology cycle



3f - How is DEISA enhancing HPC services in Europe?

- **Running larger parallel applications** in individual sites, by a cooperative reorganization of the global computational workload on the whole infrastructure, or by the operation of the **job migration service** inside the AIX super-cluster.
- Enabling **workflow applications** with UNICORE (complex applications that are pipelined over several computing platforms)
- Enabling coupled multiphysics Grid applications (when it makes sense)
- Providing a **global data management** service whose primordial objectives are:
 - Integrating distributed data with distributed computing platforms
 - **Enabling efficient, high performance access to remote datasets** (with Global File Systems and stripped GridFTP). We believe that this service is critical for the operation of (possible) future European petascale systems
 - Integrating hierarchical storage management and databases in the supercomputing Grid.
- **Deploying portals** as a way to hide complex environments to new users communities, and to interoperate with another existing grid infrastructures.

4 – Infrastructure organization

- **Description of the project organization and roadmap**

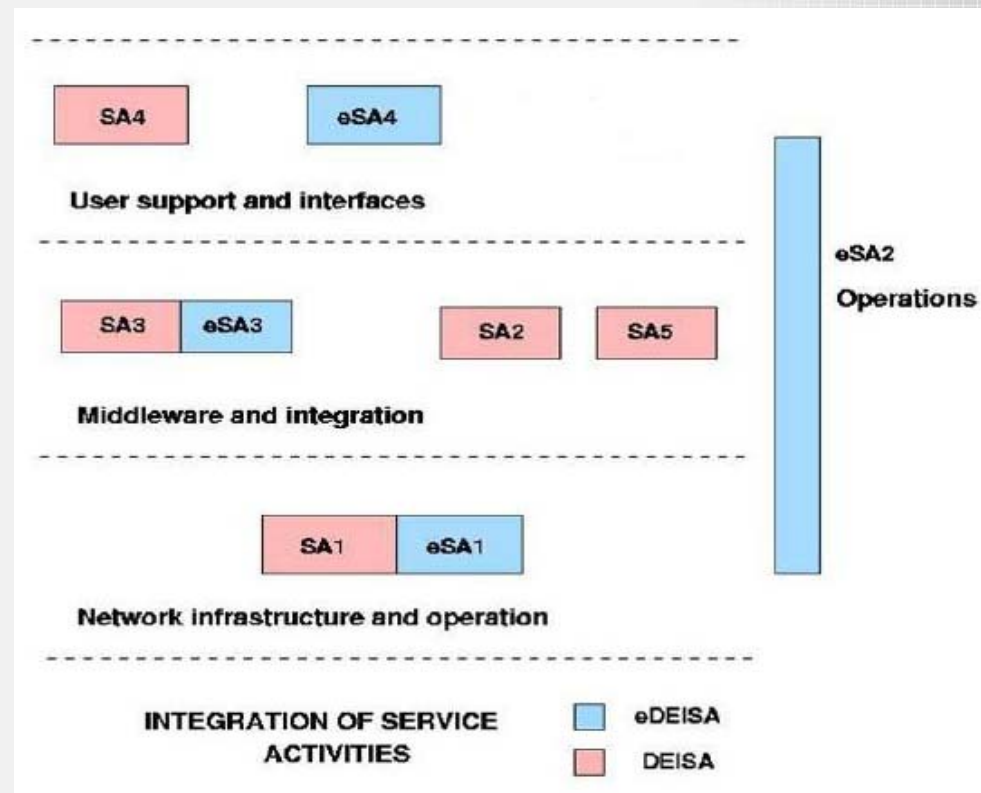
4a - Project organization

- System Integration, deployment and operation of the infrastructure:

- SA1 Networking
- SA2 Global File Systems
- SA3 + eSA3 Middleware and Ressource Management
- SA4 User Support
- eSA4 Applications Enabling and Benchmarking
- SA5 Security
- eSA2 Operations
- JRA7 Heterogeneous resource management (JRA7)

- Applications

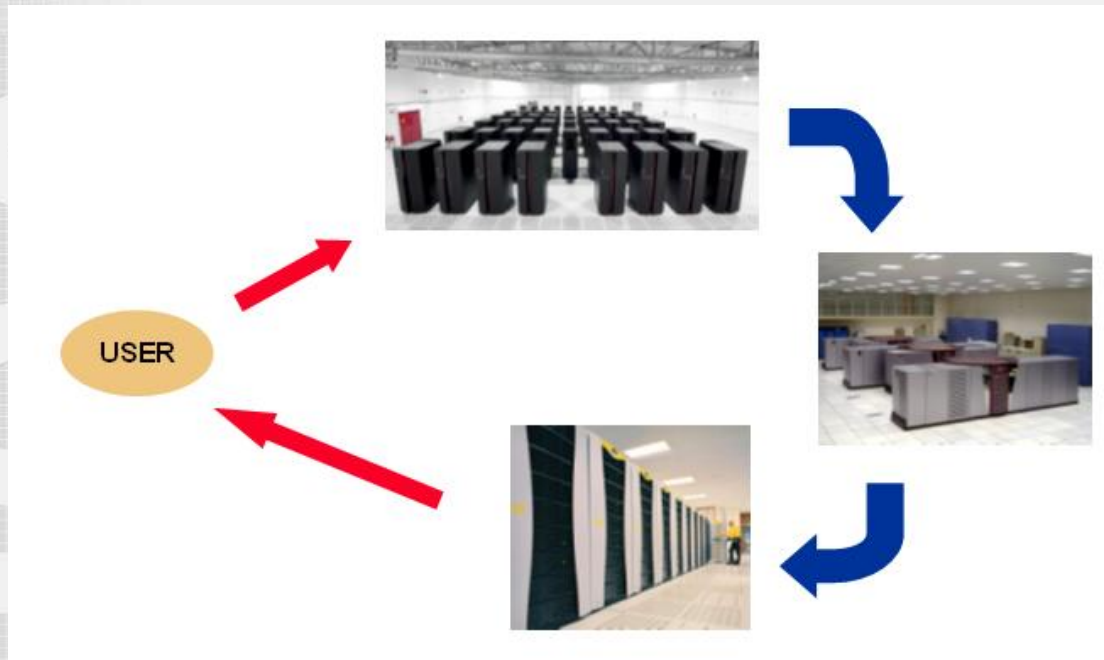
- Scientific JRAs (JRA1 to JRA6)
- The Applications Task Force
- The DEISA Extreme Computing Initiative



5 – Running the infrastructure: services

- **Description of the basic DEISA supercomputing services:**
 - **UNICORE**
 - **Global File Systems**
 - **Job Migration**
 - **Co-scheduling**
 - **Fast file transfer**
 - **Common production Environment**

5a: Workflow simulations using UNICORE



UNICORE supports complex simulations that are pipelined over several heterogeneous platforms (workflows).

UNICORE handles workflows as a unique job and transparently moves the output – input data along the pipeline.

UNICORE clients that monitor the application can run in laptops.

UNICORE has a user friendly graphical interface. DEISA has developed a command line interface for UNICORE.

**UNICORE infrastructure including all sites has full production status.
It has proven to be very stable during the last few months.**

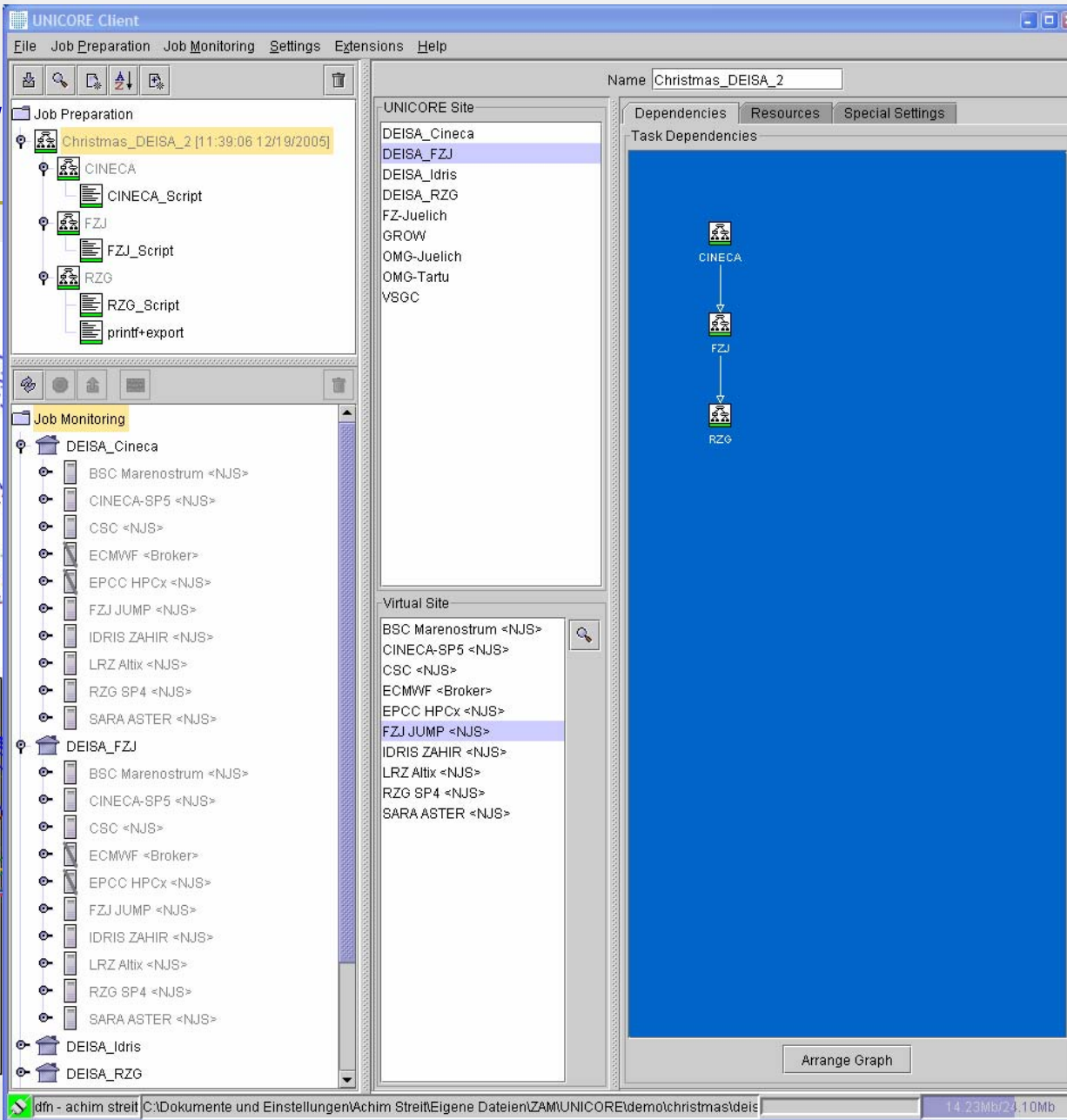
5a: W

FZJ users

DEISA
FZJ
gateway

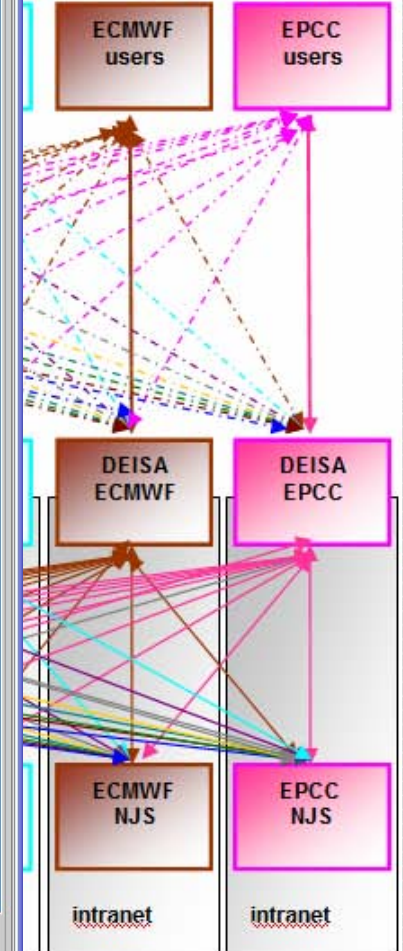
FZJ NJS

intranet

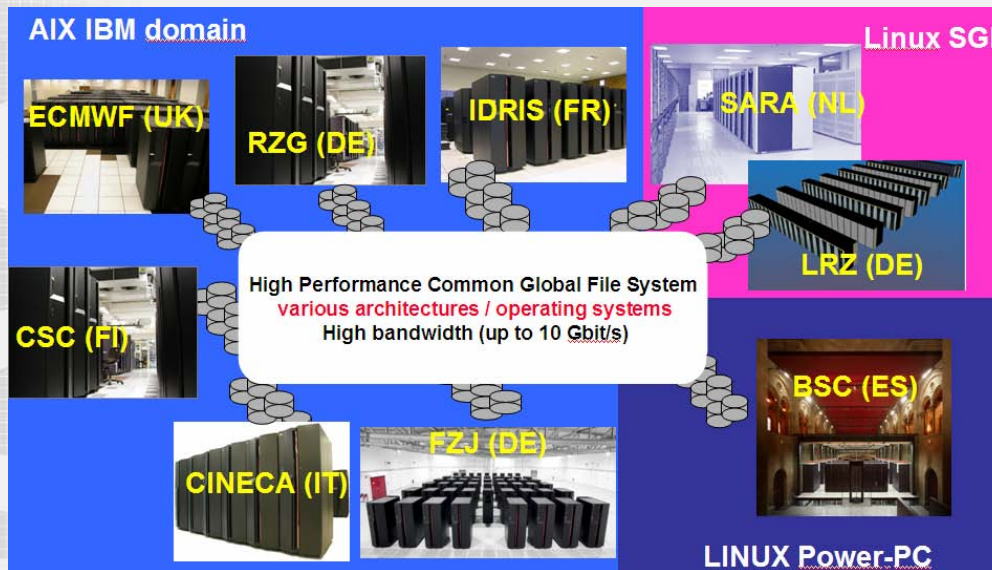


The screenshot shows the UNICORE Client interface with the following sections:

- Job Preparation:** Shows a tree view for job 'Christmas_DEISA_2 [11:39:06 12/19/2005]' with sub-items for CINECA, FZJ, and RZG, each with associated script files.
- Job Monitoring:** Lists various sites and resources such as BSC Marenostrum, CINECA-SP5, CSC, ECMWF, EPCC HPCx, FZJ JUMP, IDRIS ZAHIR, LRZ Altix, RZG SP4, SARA ASTER, DEISA_FZJ, DEISA_Idris, and DEISA_RZG.
- UNICORE Site:** Lists sites including DEISA_Cineca, DEISA_FZJ, DEISA_Idris, DEISA_RZG, FZ-Juelich, GROW, OMG-Juelich, OMG-Tartu, and VSGC.
- Task Dependencies:** A graph showing a vertical dependency chain: CINECA → FZJ → RZG.
- Virtual Site:** Lists virtual sites like BSC Marenostrum, CINECA-SP5, CSC, ECMWF, EPCC HPCx, FZJ JUMP, IDRIS ZAHIR, LRZ Altix, RZG SP4, and SARA ASTER.



5a - Global File Systems



Sophisticated software environment, necessary to provide single system image if a clustered computing platform.

They provide global data management. Data in the GFS is “symmetric” with respect to all computing nodes.

GFS encapsulate sophisticated distributed computing and Grid technologies. Applications do not need to be modified to benefit from GFS services.

IBM GPFS is a Global File System **that can be deployed over a WAN** (distributed cluster). Users see remote files as if they were local files.

The “P” (for parallel) allows GPFS to use multiple TCP streams to move large chunks of data, using the full network bandwidth. This enables high performance access to remote data.

5b - Global File System usages

- Basic usage is **DATA SHARING**: data placed in DEISA GPFS can be accessed with comparable (high) performance from any computing platform. This enables the deployment of cooperative working environments for trans-national scientific collaborations.
- DEISA uses GPFS to enable a **JOB MIGRATION** service inside the AIX super-cluster. If a user places code and data in DEISA GPFS and selects the Common Production Environment, then the multi-cluster function in Load leveler allows the transparent migration of the job to another similar computing platform.
- GPFS is essentially a very efficient network file system. Remote file systems are mounted on local file systems without performance penalty. I/O performance is mainly controlled by local file system performance.
- Remote files appear as local files to the user, but remote the data is implicitly moved to the local system. GPFS can be used to efficiently move huge data sets across computing platforms

Global File System Interoperability demo during Supercomputing Conference 2005 in Seattle

American and European supercomputing infrastructures linked:
bridging communities with scalable, wide-area global file systems

TeraGrid Sites



DEISA Sites



5c - Other basic services

- **Job migration inside the AIX super-cluster.** Based on LoadLeveler Multi-Cluster, it allows system administrators to reroute jobs to other sites, in a way transparent for the end users. Used to move away simple jobs of « implicit users » to make place for a bigger application in a site. **Full production status.**
- **Co-allocation.** We are starting to prepare a first generation co-allocation service on the full heterogeneous infrastructure, using LSF Multi-cluster. Important for coupled Grid applications and for data movement. **Service in development phase, prototype expected in 6-9 months**
- **Remote I/O and fast data transfers with GridFTP.** See next transparency
- **Integrating hierarchical data management and databases in the supercomputing Grid.** **In progress.**
- **Deploying and monitoring a Common Production Environment.** **Operational over the whole infrastructure.**

6. Running the Infrastructure: applications

- **Actions to enable leading computational science in Europe**
 - Initial program: a number of Joint Research Activities integrated in the project from the start.
 - Moving towards « exceptional users » : the DEISA Extreme Computing Initiative launched in 2005.
 - The DEISA Life Sciences Portal (planned)

JRA1 - Materials Sciences



Alexandria - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

DEISA

Alexandria
JRA1 & JRA3 Applications' Portal

DISTRIBUTED EUROPEAN INFRASTRUCTURE FOR SUPERCOMPUTING APPLICATIONS

Home | Science Areas | SA Components | Tools | Links | Contact | Help

Welcome

to the DEISA's Application Portal of the Joint Research Activities (JRAs) 1 and 3

- Material Sciences (JRA1)**

Physical, chemical, and biological processes for many problems in computational physics, biology, and materials sciences span length and time scales of many orders of magnitude. For example, on the microscopic level, the typical bond distance between atoms is of the order of Angstroms (the lattice constant). More ...
- Plasma Physics (JRA3)**

Research on magnetic confinement fusion has undergone large changes during the last decade, moving away from the semi-empirical, predominantly experiment-driven approach to one accompanied and supported in all areas by first-principle based modelling. This development has been particularly dramatic in the area of turbulence. More ...

General DEISA news

JRA 1 & JRA3 Applications' Portal available
Wed., 25 Jan 2006 14:00:00 CET
Starting Feb 01, DEISA's JRA1 and JRA3 will become available to the all DEISA extreme computing initiative users. Use the link to register for portal use. (active Feb 01) More ...

JRA 1 & JRA3 publish news feed on their Applications' Portal.
Tue., 24 Jan 2006 11:06:42 CET
Starting Feb 01, DEISA's JRA1 and JRA3 publish a news feed using the RSS technology. More ...

Portal v1.08 hosted by RZG - Max Planck Gesellschaft powered by Cocoon

Full support of the CPMD application within DEISA

NEC SX-8 CPMD optimization

Support of QUICKSTEP code in DEISA

Implementation of a portal for Materials Sciences and Plasma physics

JRA2 - Cosmology: objectives JRA3: Plasma Physics

- to avail the Virgo Consortium of the most advanced features of Grid computing by porting their production applications
 - GADGET and FLASH
- to make an effective use of the DEISA infrastructure
- to lay the foundations of a Theoretical Virtual Observatory (VirtU).
 - JRA2 funded 50/50 by DEISA and VirtU
- EPCC works in close partnership with the Virgo Consortium
 - JRA2 managed jointly by Dr Gavin Pringle (EPCC/DEISA) and Prof. Carlos Frenk (co-PI of both Virgo and VirtU)
 - work progressed after gathering clear user requirements from Virgo Consortium.
 - requirements and results published as public DEISA deliverables.

JRA3 - Plasma Physics

TORB code (turbulent transport in plasmas) related activities

A trans-national European collaboration for developing the TORB code for stellarator configurations has been established between CIEMAT, Spain and IPP, Germany.

After the TORB code was ported to the MareNostrum supercomputer of BSC, simulations were continued at BSC.

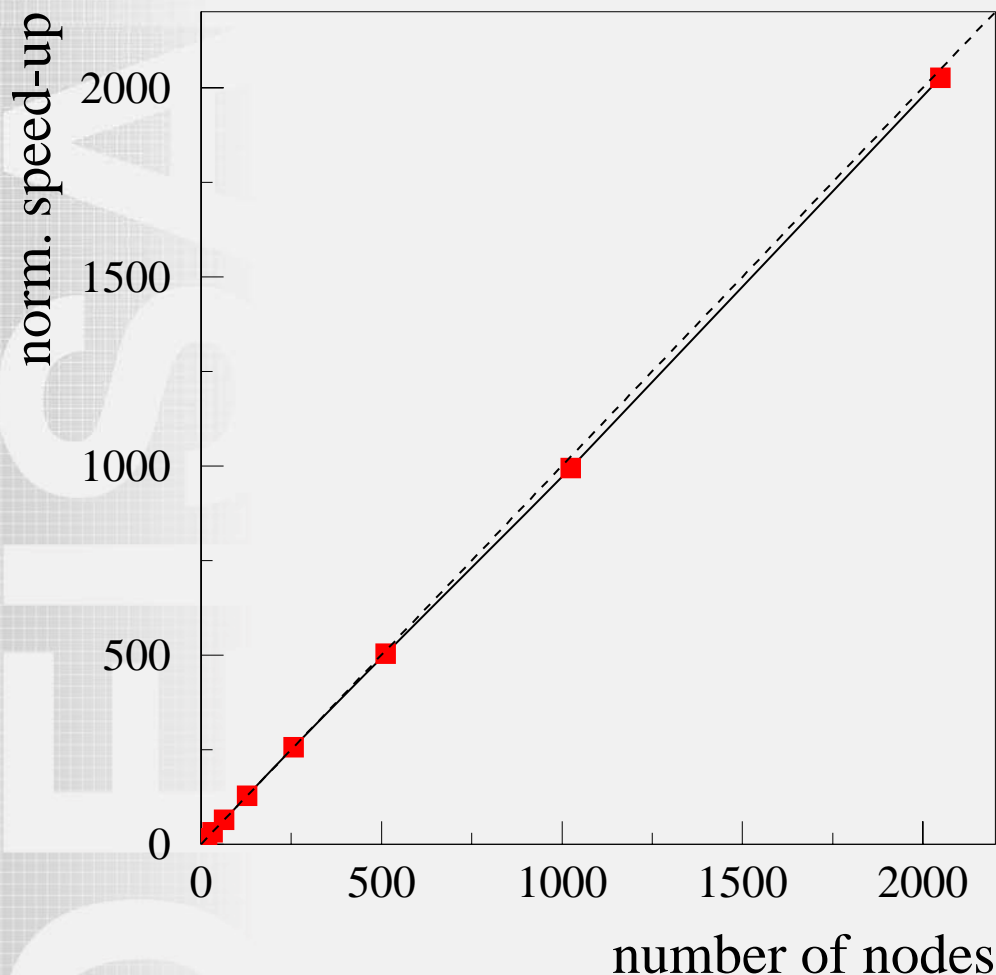
The results have been presented at the IAEA Technical Meeting on Innovative Concepts and Theory of Stellarators, Madrid, October 10-11, 2005.

For the DEISA-TeraGrid interoperability demonstration during the Supercomputing Conference 2005 in Seattle, TORB was a featured DEISA application

Paper :

R. Hatzky: Domain Cloning for a Particle-in-Cell (PIC) Code on a Cluster of Symmetric-Multiprocessor (SMP) Computers, Parallel Computing, in press

Finalizing work on TORB code



TORB Hyperscalability on MareNostrum at BSC

Result with 4096 processors:

Speed-up = 4000

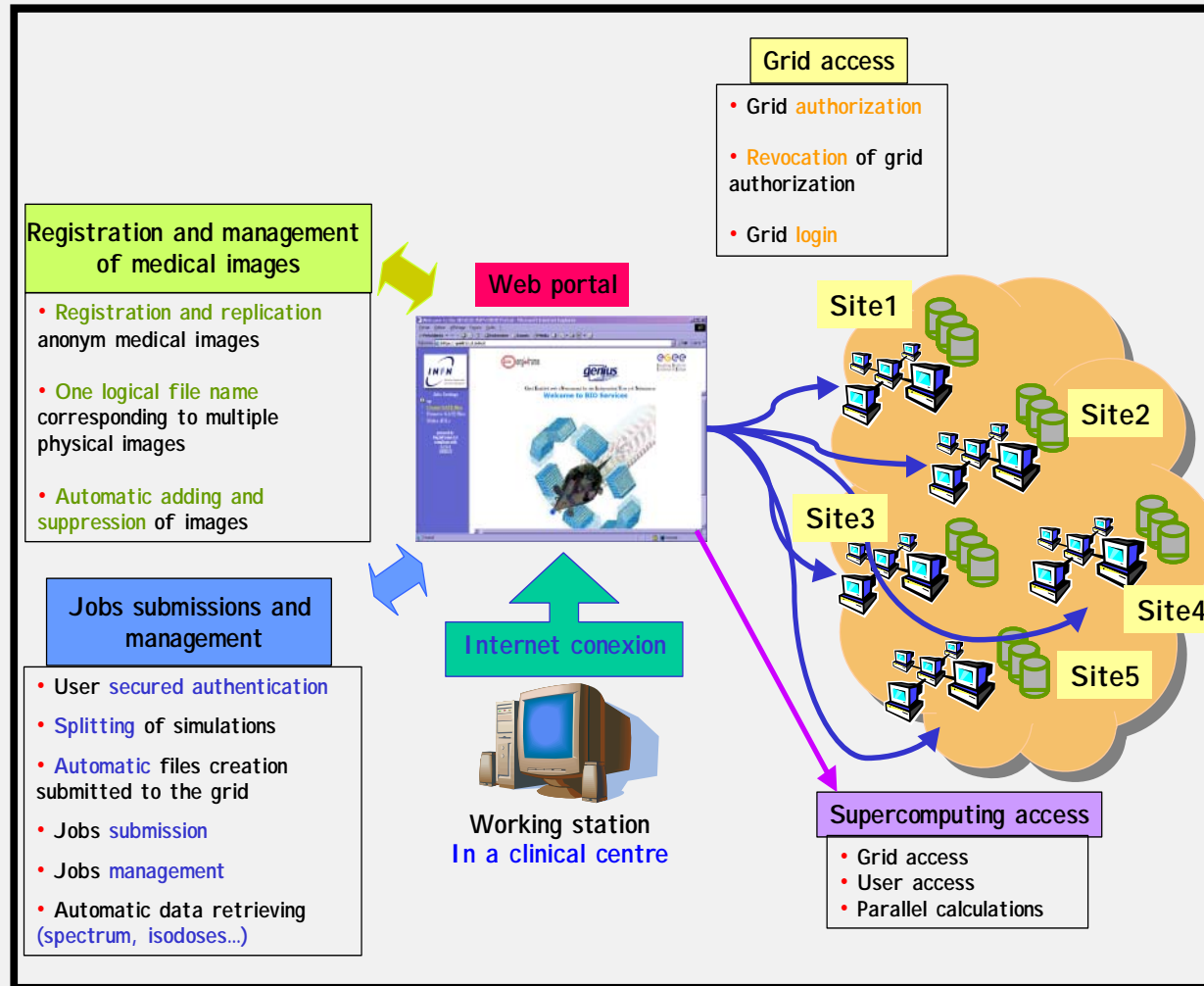
Parallel efficiency = 99%

Sustained

performance = 2.5 TF

(2048 nodes = 4096 procs)

JRA4 - Radiation Therapy planning (joint application with EGEE)

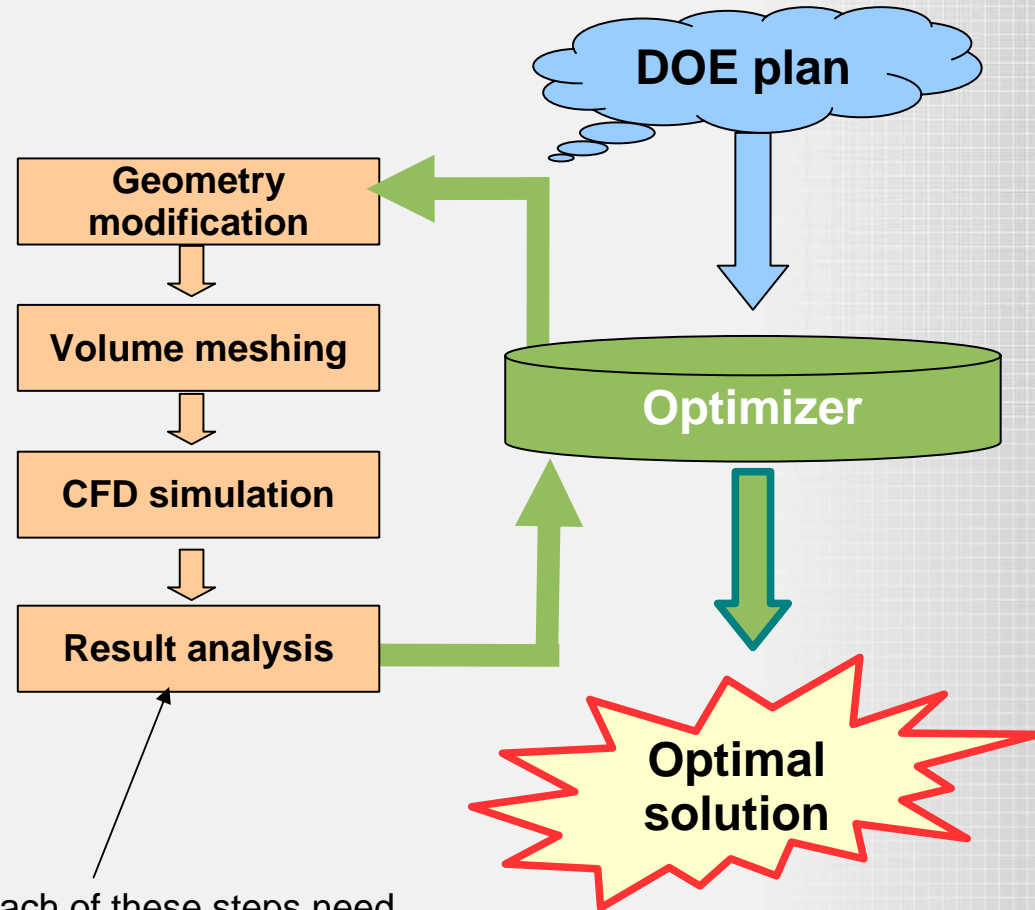
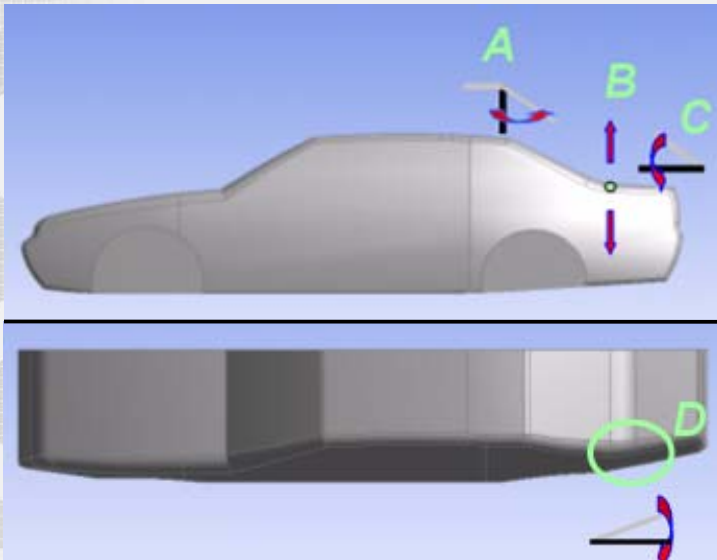


JRA5 - Industrial CFD and CAA app.

Objectives:

- Demonstrate the use of commercial CFD and CAA codes on the DEISA infrastructure
- Exploit DEISA capabilities for the use of commercial codes:
 - code optimisation
 - licensing issues
 - advanced scheduling integration
- Raise the limit of industrial simulations capabilities a step forward:
 - increase the complexity of simulations
 - increase of engineering relevance of simulations
- constraint: time to get results
- Understand how to set up commercial codes ASP service into the DEISA infrastructure

JRA5 - Aerodynamic shape optimization



- 4 parameters to be optimized
- cubic face centered DOE
- 25 cases+16 extra cases for error estim.
- polynomial response function
- 70 hours wall clock time on 64 cpus

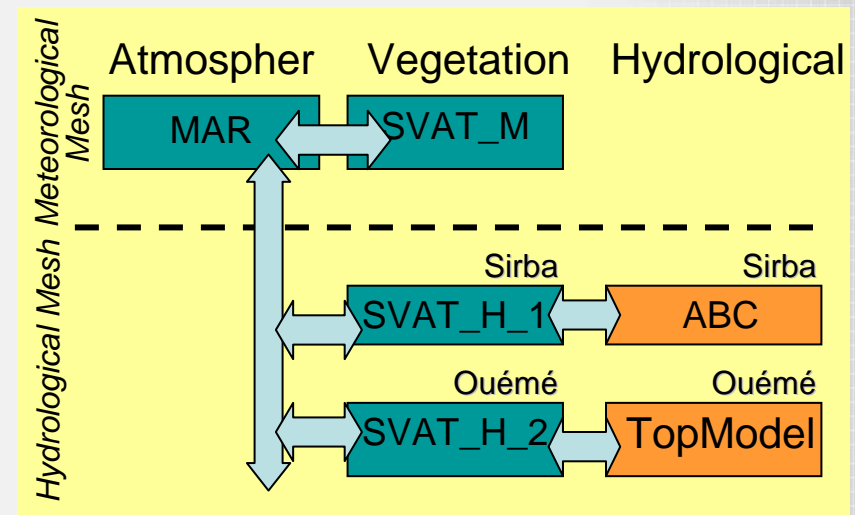
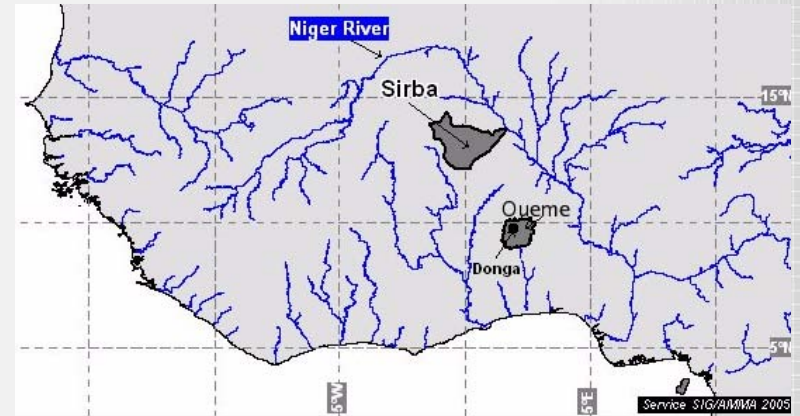
Each of these steps need to be fully automated and controlled by the optimizer

JRA6 - Coupled applications overview

- **Objectives**
 - To adapt coupling projects to the DEISA infrastructure and to improve the parallelism:
 - at the code level
 - at the coupling level (asynchronous models)
 - To submit coupling projects to the DECI
- **The initial projects (PM 1 - PM 18)**
 - 3 projects running today on heterogeneous platforms
- **The second set of projects (PM 19 – PM 36)**
 - Chosen among HLRS and IDRIS projects.

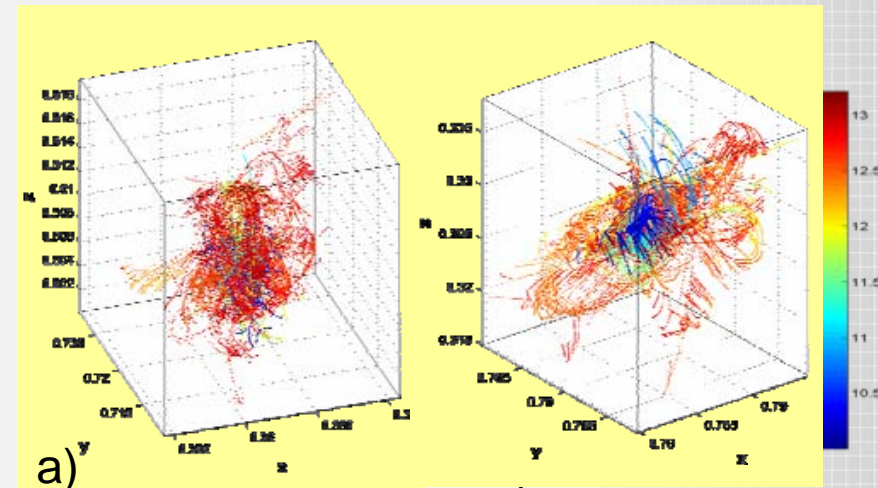
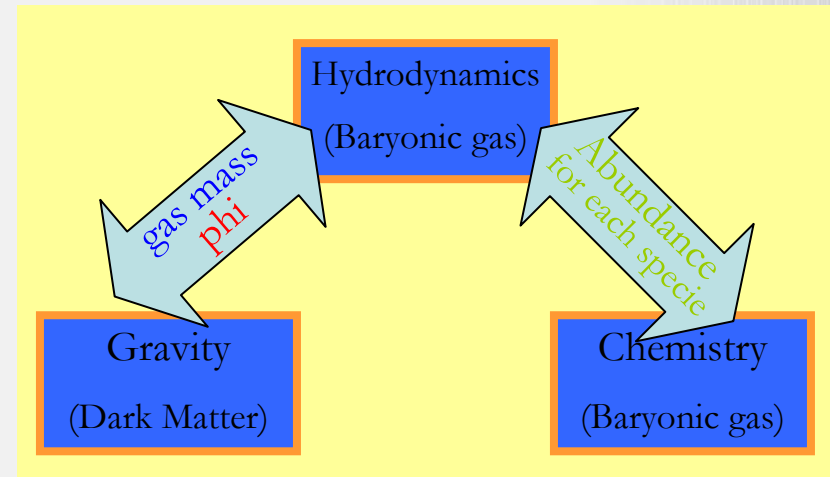
Initial projects (1)

- Environmental project:
 - Study the impact of water cycles of the hydrological and vegetation models on climate models
 - Coupling area in West Africa
 - Best performances with a vector and scalar platform
 - Improve extensibility of the architecture and the coupling part
 - AMMA project, PhD thesis, 2 publications and 2 communications



Initial projects (2)

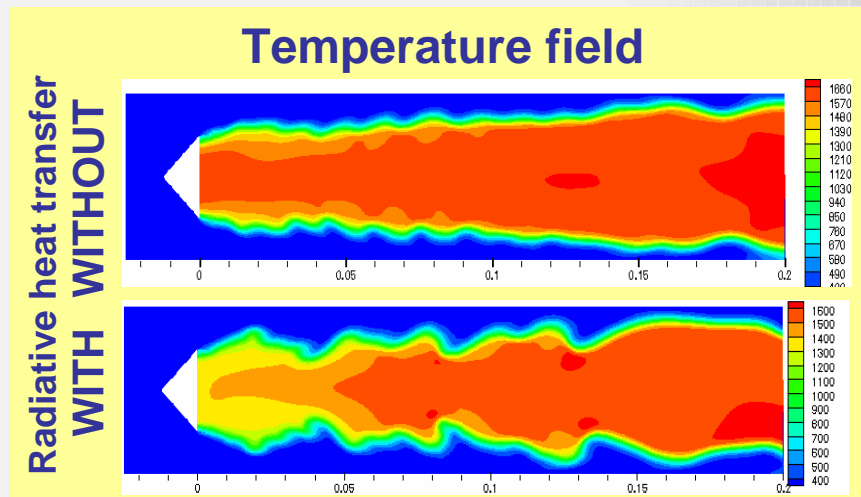
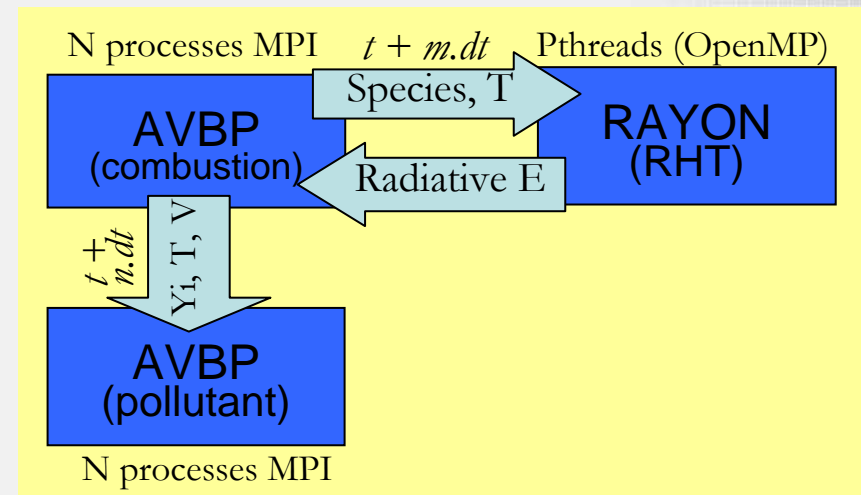
- **Cosmological project:**
 - Study galaxy formation in cosmology
 - Physics / module: Gravitation, Hydrodynamics, Chemistry
 - Best performance on heterogeneous platforms
 - Load balancing issue and improvement of the coupling part
 - Proposed to DECI
 - HORIZON project, 1 publication, 2 communications



Initial projects (3)

- **Combustion / Radiation project**

- Study the impact of radiative heat transfer (RHT) on the combustion process (2D)
- Couple combustion (AVBP), the RHT (Rayon) codes and the pollutant formation (AVBP)
- Parallelization of the Rayon code and improvement of the coupling part
- Load balancing issue
- 3D extension proposed to DECI and accepted
- PhD thesis, 2 communications in international congresses



6b - DECI: enabling leading computational science

- The basic service providing model for scientific users is the **DEISA Extreme Computing Initiative** (see www.deisa.org)
- Identification, deployment and operation of a number of « flagship » applications requiring the infrastructure services, in selected areas of science and technology.
- European Call for proposals in May-June every year. Applications are selected on the basis of scientific excellence, innovation potential and relevance criteria, with the collaboration of the HPC national evaluation committees.
- There are 29 projects in operation in 2005 – 2006 (see www.deisa.org)
- Supported by the **Applications Task Force** (ATASKF), whose objective is to enable and deploy the Extreme Computing applications. The activities of the ATASKF are focused on:
 - Hyperscaling of huge parallel applications, data oriented applications
 - Workflows and coupled applications
 - **Production of an European Benchmark Suite for HPC systems**

6c - The Life Sciences Portal

- The objective is to extending the outreach of the supercomputing infrastructure by reaching new users communities that have already structured their applications strategies around small discipline oriented grid infrastructures with discipline specific tools.
- The strategy is to connect supercomputer environments as « backend » resources to existing discipline oriented infrastructures.
- Community allocations will enable the access of external anonymous users.
- To move in this direction, DEISA plans to deploy in 2006-2007a portal to a European supercomputing service for bio-informatics and Life Sciences. This is a discipline in which the need of HPC is strongly emerging in some domains.
 - Critical domain applications are ported to the most adapted supercomputer of the DEISA environment
 - Shared data repositories are hosted by GPFS services
 - A DEISA portal will be deployed, but interoperability with existing portals will be searched
- **This ASP (Application Service Provider) model is potentially well adapted to external corporate users.**

6d - Collaboration with other projects

- **DEISA and EGEE have complementary objectives and strategies. Close collaboration is mandatory. Planned joint meetings to decide on cooperation strategy. The main interest is in applications that require access to both infrastructures.**
- **TeraGrid and DEISA have common visions on the deployment and operation of supercomputing Grids.**
- **Main subjects of TG-DEISA collaboration: GPFS, INCA (the software monitoring tool for the CPE), stripped GridFTP.**
- **DEISA is deeply engaged in OGF/GGF's Grid Interoperability Now (GIN) initiative involving all the Major Grid Projects in the world.**

8 - Conclusions

- DEISA aims at deploying a **sustained and persistent**, basic European infrastructure for general purpose high performance computing.
- **We expect services and existing synergies to be persistent.** We do not claim persistency of the current organizational model. The DEISA Consortium is ready to adapt to the new FP7 strategies and establish a roadmap incorporating cooperation or merging with new HPC initiatives.
- Our next challenge is **establishing an efficient organization embracing all relevant HPC organizations in Europe.**
- One possibility is moving, whenever possible, from a consortium of HPC sites to a consortium of HPC national organizations.
- Interfaced with the other grid-enabled complementary infrastructures, DEISA expects to continue to contribute to a global European infrastructure for science and technology
- ***Integrating leading supercomputing platforms with Grid technologies and reinforcing capability with shared petascale systems is needed to open the way to new research dimensions in Europe.***