Distributed European Infrastructure for Supercomputing Applications
DEISA
Towards a European HPC Ecosystem

Wolfgang Gentzsch, Deisa2 & Duke
Stefan Heinzel, PI Deisa2, RZ Garching
Victor Alessandrini, PI Deisa1, IDRIS
Paolo Malfetti, CINECA
Michael Rambadt, FZ Juelich
Content

Today’s HPC Scenario, and what the HPC users really want

DEISA and DEISA-2, Vision, Objectives, Requirements

The European Tier0 / Tier1 Centers

DEISA Partners and Services

Core Technologies

Applications

DECI
Today’s HPC Scenario

- Single HPC site  -  HPC Silos  -  each with specific sign-on
- Local access - No resource choice - Long waits (queues)
- Application and data has to be in the same place
- Different sites mean different admin domains
- Collaboration beyond one site a challenge
- Ecological boundaries
HPC Follows General IT Transition

Old World
- Static
- Silo
- Physical
- Manual
- Application

New World
- Dynamic
- Shared
- Virtual
- Automated
- Service

Transitioning from Silo Oriented Architecture to Service Oriented Architecture

Courtesy Mark Linesch, former OGF

HPCS 2008 Nicosia

Wolfgang Gentzsch
What HPC Users Really Want

-- HPC users are conservative, standard access methods are preferred, no interest in complicated middleware stacks.
-- Comfortable Data Access from the desktop and among HPC centres is the key for satisfied users and for the future European HPC Infrastructures!

Global Login

*Personal Login in each system, i.e. single “European” username, uniform access*

Comfortable Data Access

*Global, fast and comfortable access to their data, across the centres*

Common Production Environment

*Not need for an identical but for an equivalent HPC software stack*

Global Help Desk

*One central point of contact and as local as possible*

Application Support

Need help in scalability, performance and adaptation to different architectures
DEISA2 – The Virtual European HPC Center
Operation - Technology - Application

Objectives:
Enhancing the existing distributed European HPC environment (DEISA) to a turnkey operational infrastructure
Advancing computational sciences in Europe by supporting user communities and extreme computing projects
Enhancing service provision by offering a complete variety of options of interaction with computational resources
Integration of Tier1 and Tier0 centers (currently Terascale & Petascale)
   The Petascale Systems need a transparent access from and into the national data repositories
Bridging worldwide HPC and Grid projects

HPCS 2008 Nicosia
Wolfgang Gentzsch
DEISA2 Vision: European HPC Ecosystem

DEISA2 - The Infrastructure for the European HPC Ecosystem

Deep operational and technological integration of European HPC (T-0 and T-1) centres and systems providing efficient seamless access to shared HPC resources and large data repositories designing and approving an operational model for a large European Virtual HPC Centre.

Providing scientists access to a large distributed HPC environment via integrated services.

DEISA is paving the way to the efficient operation of the T-0 and T-1 ecosystem.

Enhancing the performance

PRACE

Building a world-class pan-European High Performance HPC Ecosystem which is operated under the umbrella of a European Legal Entity adopting operational and technological concepts and services designed and approved by DEISA2.
Tier0 / Tier1 Top Layer of the HPC Ecosystem

**T0** : future shared petascale European systems

**T1** : leading national supercomputing systems

**PRACE**

Designing an infrastructure that will enable the operation of shared petascale European systems

*Enhancing performance in selected sites and providing wide access to shared systems*

---

In the DEISA-2 environment, scientists will naturally have access to several distributed platforms, and shared systems will have to access remote data repositories. We will be in a situation similar to TeraGrid.

**DEISA-2** : strong integration of T0 and T1 systems (automatically provides wide, seamless and efficient access to shared systems and data repositories)

*The DEISA-1 services have been tailored for this mode of operation. There is a positive feedback between the two orthogonal lines of action:*

  - **DEISA is paving the way to the efficient operation of T0 systems.**
  - **T0 systems will drive the massive adoption of the DEISA services.**
Tier0 / Tier1 Centers
Are there implications for the services?

Main difference between T0 and T1 centers: policy and usage models!

- T1 centers can evolve to T0 for strategic/political reasons
- T0 machines automatically degrade to T1 level by aging

T0 Centers
- Leadership-class European systems in competition to the leading systems worldwide, cyclically renewed
- Governance structure to be provided by European organization

T1 Centers
- Leading national Centers, cyclically renewed, optionally surpassing the performance of older T0 machines
- National Governance structure

The Services have to be the same in T0/T1
- Because of the change of the status of the systems, over time
- For user transparency of the different systems
- (Only visible: Some services could have different flavors for T0 and T1)
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.
Objectives of the European DEISA1 Project

Enabling Europe’s terascale science by integrating Europe’s most powerful HPC systems

Only Criterion for Success: Enabling scientific discovery across a broad spectrum of science and technology

DEISA is a European Supercomputing Service built on top of existing national HPC 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)
Objectives of the European DEISA2 Project

Enhancing the existing distributed European HPC (DEISA1) environment towards a turnkey operational infrastructure

Advancing computational sciences in Europe by supporting user communities and extreme computing projects

Enhancing service provision by offering a complete variety of options of interaction with computational resources

Integration of T1 and T0 centers

The Petascale Systems need a transparent access from and into the national data repositories

Bridging worldwide HPC and Grid projects
Basic Requirements for the DEISA Research Infrastructure

Fast Deployment of **persistent, production quality**, grid empowered supercomputing infrastructure with continental scope.

The European supercomputing service built on top of existing national services requires **reliability** and non disruptive behavior.

**User Transparency** : users should not be disturbed by technology overhead. Grid technologies should work seamlessly in the background.

**Application Transparency** : applications are part of the corporate wealth of virtual organizations, and they should be portable and independent (as much as possible) of the underlying Grid technologies.

**Top-Bottom Approach**: technology choices follow from the business and operational models of the virtual organization. DEISA technology choices are pragmatic and fully open. There is practically no « DEISA specific middleware ».

Focus on **HPC**
In DEISA, no dependency of any site from others should be introduced. This means that...

→ the policies chosen for DEISA must not interfere with local site policies, this impose the need for a homogenisation process;

→ the design architecture chosen for DEISA should be implementable by each site in a different manner without compromising the whole design or the service efficiency.
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 multi-physics Grid applications (when it makes sense)

Providing a global data management service whose main 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.
DEISA Partners

DEISA1: May 1st, 2004 – April 30th, 2008
DEISA2: May 1st, 2008 – April 30th, 2011
DEISA2 Partners

- BSC: Barcelona Supercomputing Centre, Spain
- CINECA: Consortio Interuniversitario per il Calcolo Automatico, Italy
- CSC: Finnish Information Technology Centre for Science, Finland
- EPCC: University of Edinburgh and CCLRC, UK
- ECMWF: European Centre for Medium-Range Weather Forecast, UK (int)
- FZJ: Research Centre Juelich, Germany
- HLRS: High Performance Computing Centre Stuttgart, Germany
- IDRIS: Institut du Développement et des Ressources en Informatique Scientifique - CNRS, France
- LRZ: Leibniz Rechenzentrum Munich, Germany
- RZG: Rechenzentrum Garching of the Max Planck Society, Germany
- SARA: Dutch National High Performance Computing, Netherlands
- KTH: Kungliga Tekniska Högskolan, Sweden
- CSCS: Swiss National Supercomputing Centre, Switzerland
- JSCC: Joint Supercomputer Center of the Russian Academy of Sciences, Russia
The DEISA supercomputing environment
(21.900 processors and 145 Tf in 2006, more than 190 Tf in 2007)

IBM AIX Super-cluster

FZJ-Julich, 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, Linux cluster (2.7 teraflops) moving to SGI ALTIX system (5120 processors and **33 teraflops peak in 2006, 70 teraflops peak in 2007**)

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**
The DEISA Supercomputing Grid

- AIX distributed super-cluster
- Vector systems (NEC, …)
- Linux systems (SGI, IBM, …)

DEISA
DEISA Partner Networks
<table>
<thead>
<tr>
<th>2005</th>
<th>2006</th>
<th>2007</th>
<th>2008</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Gbps network connectivity</td>
<td>Enlargement of WAN shared file system</td>
<td>10 Gbps network connectivity</td>
<td>10 Gbps Unified AAA</td>
</tr>
<tr>
<td>Common DEISA policies</td>
<td>Enhanced job rerouting</td>
<td>Unified accounting</td>
<td>Unified data management strategy (WAN shared fs + archive + data staging + data transfer)</td>
</tr>
<tr>
<td>Unified Auth</td>
<td>Enlargement of workflow management with data staging tools</td>
<td>Fully fledged WAN shared filesystem</td>
<td></td>
</tr>
<tr>
<td>WAN shared filesystem</td>
<td>CLI access</td>
<td>2nd common production environment</td>
<td></td>
</tr>
<tr>
<td>Job rerouting</td>
<td>Single monitoring system</td>
<td>Portal access for life science community</td>
<td>Enriched middleware infrastructure for interoperability</td>
</tr>
<tr>
<td>1st common production environment</td>
<td>Data staging tools</td>
<td></td>
<td>Multiple way to access DEISA</td>
</tr>
<tr>
<td>Workflow management</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
DEISA Services

- Multiple ways to access
- Workflow management
- Common production environment
- Presentation layer

- Single monitor system
- Job rerouting
- Co-reservation and co-allocation
- Job management layer and monitor

- Data staging tools
- Data transfer tools
- WAN shared filesystem
- Data management layer

- Unified AAA
- DEISA Sites
- Network connectivity
- Network and AAA layers
DEISA
a service-oriented production (RAS)
e-Infrastructure

Multiple ways to access
Workflow managem.
Common production environm.

Sites
Unified
Unified

Distributed
European
Infrastructure for
Supercomputing
Applications

Unified AAA
DEISA Sites
Network connectivity
The DEISA Service Level Hierarchy Model

DEISA identifies Core Services (Level1)
and Additional Services (Level2 and Level3)

Core Services (Level1)
Key technologies with the following characteristics:
Essential, robust and reliable, pure HPC oriented, generally accepted

Additional Services (Level2 and Level3)
Technologies with the following characteristics:

Level2
Additional ways to Access HPC systems, Workflow services with data staging
Specialized needs of Virtual Communities (i.e Climate, Fusion,…)
Interoperability with other HPC infrastructures as Teragrid, EGEE, etc.

Level3
"Nice to have" with potential for Level2
Not necessarily production quality
The DEISA Core Infrastructure and Services

Dedicated high speed network (WP3)

Common AAA (WP3 and WP4)
  - Single sign on
  - Accounting/budgeting

Global data management (WP3 and WP4)
  - High performance remote I/O and data sharing with global file systems
  - High performance transfers of large data sets

User Operational infrastructure (WP3 and WP6)
  - Distributed Common Production Environment (DCPE)
  - Job management service
  - Common user support and help desk

System Operational infrastructure (WP3 and WP4)
  - Common monitoring and information systems
  - Common system operation

Global Application Support (WP5, WP7, WP9)
Motivation: Why UNICORE?

- Scientists have to use huge computational and storage resources...
Motivation: Why **UNICORE**?

Supercomputers are managed by Resource Management Systems (RMSs) that handle the scheduling.

But: There are many RMSs available →

There are many proprietary ways of job submission
- IBM Loadleveler → **lsl**
  submit…
- Torque Resource Manager → **qsub**…

Different job description languages
- (# of nodes, memory requirements…)
UNICORE
UNiform Interface to COmputing Resources

Started as a German Government funded Project in 1997

Available as open Source Software on www.unicore.eu

Seamless, secure access to distributed resources and data

Intuitive GUI and DESHELL

Workflow engine for
complex multi-site multi-step workflows
job monitoring

No ‘learn overhead’ if a new RMS is used

Easy installation and configuration
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.
Global File Systems

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

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

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

**IBM GPFS:** 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.
DEISA Global File System integration in 2006

(based on IBM’s GPFS)

High Performance Common Global File System

various architectures / operating systems

High bandwidth (up to 10 Gbit/s)
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 aplications and for data movement. Service in development phase, prototype status.

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.
Accessing Remote Data:  
High performance remote I/O and file transfer

Remote I/O with global file systems *implicitly* moves data across platforms (in production today)

DEISA will also deploy *explicit* high performance data movers, using a parallel implementation of GridFTP
AIX SUPER-CLUSTER, June 2006

Full production status of dedicated 1 Gb/s network
GPFS: Full production at IDRIS, RZG, FZJ, CINECA, CSC, ECMWF
JOB MIGRATION: Full production in core sites

Services:

High performance datagrid via GPFS
Access to remote files use the full available network bandwidth

Job migration across sites
Load balancing the global workflow when a huge partition is allocated to a DEISA project at one site

Common Production Environment
Material Sciences

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
Cosmology: Objectives

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

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.
TORB code (turbulent transport in plasmas) related activities

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:
Plasma Physics

Finalizing work on TORB code

TORB Hyperscalability on MaresNostrum at BSC

Result with 4096 processors:

- Speed-up = 4000
- Parallel efficiency = 99%
- Sustained performance = 2.5 TF

(2048 nodes = 4096 procs)
Industrial CFD Application

Objectives:
- Demonstrate use of commercial CFD 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 code ASP services in the DEISA infrastructure
Aerodynamic Shape Optimization

- 4 parameters to be optimized
- cubic face centered DOE
- 25 cases + 16 extra cases for error estimation
- 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.
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 (DEISA1, PM 1 - PM 18)

3 projects running today on heterogeneous platforms

The second set of projects (DEISA1, 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 publ. and 2 comms.
Initial Projects (2)

Cosmological project:

- Study galaxy formation in cosmology
- Physics / modules: Gravitation, Hydrodynamics, Chemistry
- Best performance on heterogeneous platforms
- Load balancing issue and improvement of the coupling part
- Proposed to DECI
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 comms at international congresses

Temperature field
The basic service providing model for scientific users is the Extreme Computing Initiative (see www.deisa.eu)

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 23 projects in operation in 2007, and 29 projects operated in 2006 (see www.deisa.eu)

Supported by the Applications Task Force (ATASKF), objective is to enable + deploy 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 a European Benchmark Suite for HPC systems
Enabling Science: DEISA users communities

National users communities have accounts on a given site and do not « naturally » see the whole DEISA environment.

Promotion of DEISA users is done via the Extreme Computing Initiative.

Europ. call for proposals for grand challenge simulations every year in May since 2005.

About 50 grand challenge projects supported each year since 2005.

Full information about Extreme Computing projects and reports from terminated projects can be found on the DEISA Web server: www.deisa.eu

The Extreme Computing Initiative is the current DEISA service provisioning model.
DECI – DEISA Extreme Computing Initiative

DECI Call 2005
> 50 proposals, 12 European countries involved
> 30 mio cpu-h requested

DECI Call 2006
> 40 proposals, 12 European countries involved
> 30 mio cpu-h requested

DECI Call 2007
> 60 proposals, 13 European countries involved,
> 70 mio cpu-h requested

DECI Call 2008, Deadline end of June, see www.deisa.eu
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 eInfrastructures.

Community allocations will enable the access of external anonymous users.

To move in this direction, DEISA plans to deploy in 2006-2007 a 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.**
DEISA – Summary of achievements and scientific impact (1)

- Heterogeneous supercomputing infrastructure built-up at continental scale
- Priority high speed end-to-end network
- Common global high performance file system at continental scale to greatly facilitate data management across Europe
- Harmonization of the many heterogeneous software environments through the DEISA Common Production Environment (DCPE)
- Uniform infrastructure access through the UNICORE middleware
- Job routing / re-routing across Europe with MC-LL among five AIX platforms to
  - free resources for a big job requiring close to all resources at one site
  - facilitate simultaneous usage of many platforms for independent subtasks of a big project, to accelerate project turn-around
DEISA – Summary of achievements and scientific impact (2)

• Portals for transparent access to complex supercomputing environments

• Internal services:
  - system development and deployment of continuous resource and service monitoring,
  - centralized trouble ticket system,
  - accounting facility
  - database for the management of the DECI projects

• High-level application enabling

• Setup of a European Benchmark Suite

• Enabling of cooperative extreme computing in Europe

• Grand Challenge projects performed on a regular basis
Conclusions

Production Grids are possible
But: only interesting if
  deployment of new production software offers added value
  easy usage, increased effectiveness, decreased cost, …
  integration of legacy applications

Success of the Grid Middleware depends on successful
  interaction with other components, working groups, colleagues…

Functionality is important but also support, support, support…
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 has defined clear new FP7 strategies and is establishing a roadmap incorporating cooperation or integration with new HPC initiatives.

Our next challenge is establishing an efficient organization embracing all relevant HPC organizations in Europe.

Interfaced with the other grid-enabled complementary infrastructures, DEISA expects to continue to contribute to a global European eInfrastructure 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.*