



The EU projects DEISA and PRACE - *Network requirements for HPC applications*

September, 21th 2008 | Ralph Niederberger



Overview

- **Introduction and motivation**
- **The EU projects**
 - **DEISA**
 - **PRACE**
- **Current network setup**
- **HPC application demands on networking**
- **Future infrastructure needs for HPC computing**
- **Summary**

Introduction and motivation

The EU project DEISA provides access to national HPC centres for European Scientists.

The EU project PRACE prepares the construction of a world-class European HPC service as a permanent Research Infrastructure

These supercomputers, which will provide Petaflops/s are useless, if not accessible via European-wide high speed networks, that allow scientists to interchange data between any of those systems.

The presentation will provide a short overview on both projects and will highlight some of the applications used within DEISA and PRACE.

It will describe the impact on those applications, which is highly dependent on the available network infrastructure.

Furthermore, it will envision future infrastructure needs for HPC applications desirable and those indispensable.

HPC is a “Key Technology”

Scientific research is based on the three pillars:

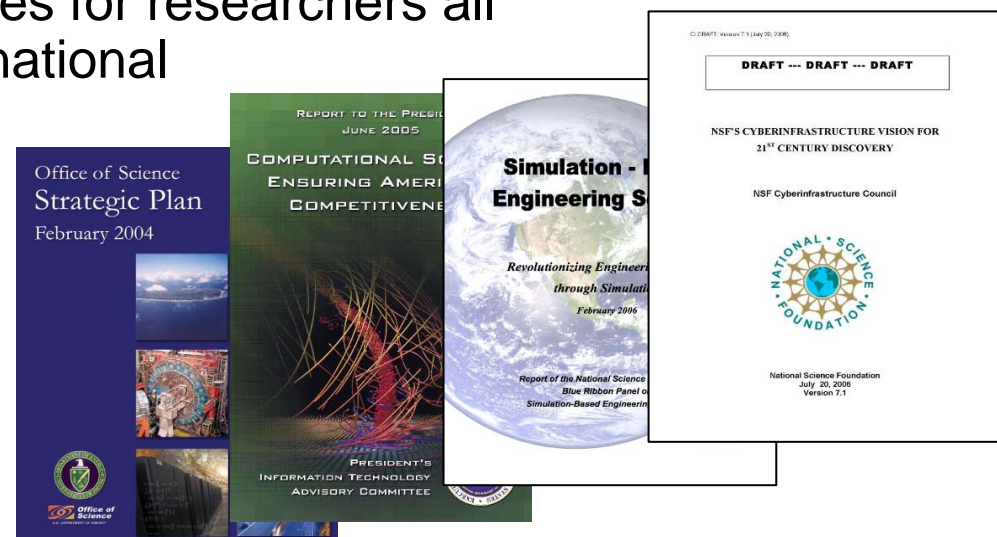
Theory, Experiment and Simulation

Supercomputers are *the* tool for solving most of those challenging problems through simulations

Access to these supercomputer systems is normally granted for a small number of scientist in a peer review process

Providing access to these services for researchers all over Europe independent of national boundaries is an ongoing challenging task

Leading industry nations such as USA and Japan acknowledge this since the 1990'ies



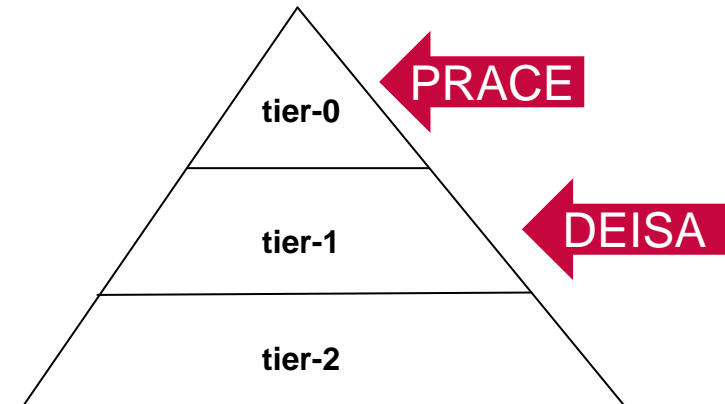
The ESFRI Vision for a European HPC service

European HPC-facilities at the top of an HPC provisioning pyramid

- Tier-0: 3-5 European Centres
- Tier-1: National Centres
- Tier-2: Regional/University Centres

Creation of a European HPC ecosystem involving all stakeholders

- HPC service providers on all tiers
- Grid Infrastructures
- Scientific and industrial user communities
- The European HPC hard- and software industry



Unlike other European Research Infrastructures:

- Tier-0 resources have to be renewed every 2-3 years
- Construction cost 200 – 400 Mio. € every 2-3 years
- Annual running cost 100 – 200 Mio. €

*1 ESFRI - European Strategy Forum on Research Infrastructures

DEISA

Distributed European Infrastructure for Supercomputer Applications

The DEISA project



DEISA Partners



DEISA: May 2004 – April 2008

eDEISA: May 2006 – May 2008 - Three new partners joined

DEISA2: May 2008 – April 2011 - Three additional associate partners joined

Basic Requirements & Goals for the DEISA Research Infrastructure

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

Reliable and non-disruptive European supercomputing service built on top of existing national services

User Transparency by seamlessly working grid technology

Application Transparency: applications should be portable and (as much as possible) independent of the underlying Grid technologies.

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

Evolution of Supercomputing Resources

DEISA partners' resources at project start in 2004:

~ 30 TF aggregated Peak performance

DEISA partners' resources in mid 2008:

Close to 1 PF aggregated Peak performance
on state-of-the art supercomputers

Cray XT4, Linux

IBM Power5, Power6, AIX / Linux

IBM BlueGene/P, Linux (frontend)

IBM PowerPC, Linux (MareNostrum)

SGI ALTIX 4700 (Itanium2 Montecito), Linux

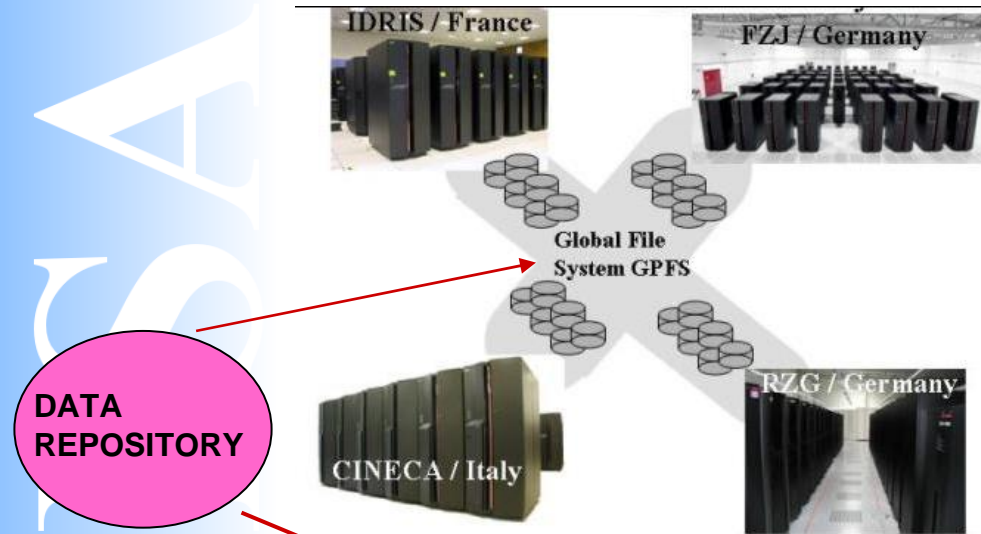
NEC SX8 vector system, Super UX

Systems interconnected with dedicated 10Gb/s

DEISA network provided by GEANT2 and NRENs

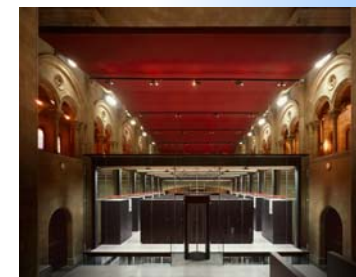
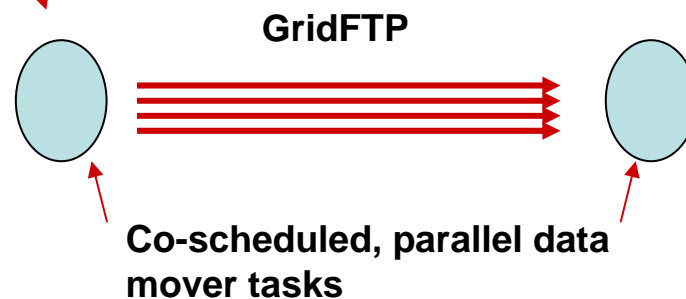
Fixed fractions of resources dedicated to DEISA usage

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





PRACE funded in part by EC under FP7 Capacities programme grant agreement INFSO-RI-211528



PRACE: European Access to HPC-Technology



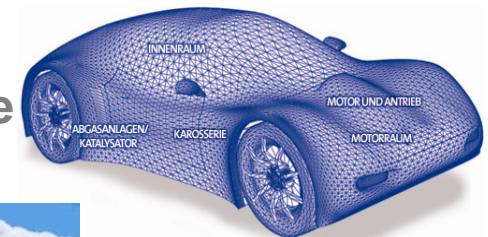
finance



aerospace



automotive



virtual power plant

European Ecosystem



PRACE – Project Facts

Objectives of the PRACE Project:

- Prepare the contracts to establish the PRACE permanent Research Infrastructure as a single Legal Entity in 2010 including governance, funding, procurement, and usage strategies.
- Perform the technical work to prepare operation of the Tier-0 systems in 2009/2010 including deployment and benchmarking of prototypes for Petaflops systems and porting, optimising, peta-scaling of applications
- Created to implement the ESFRI vision of a European HPC service




Project facts:

- Partners: 16 Legal Entities from 14 countries
- Project duration: January 2008 – December 2009
- Project budget: 20 M €, EC funding: 10 M €



PRACE – Project Consortium



New Partners - since May 2008 - of the PRACE Initiative :   



The next tasks (I/II): ... growing into a persistent Research Infrastructure

Define the legal form and governance

Secure initial and continuous funding

Prepare procurement and installation of the first Petaflops systems

Establish the peer review process for academic usage

Promote Europe wide collaboration between scientific communities using leading edge scientific simulation

Encourage new projects to increase software and simulation competence

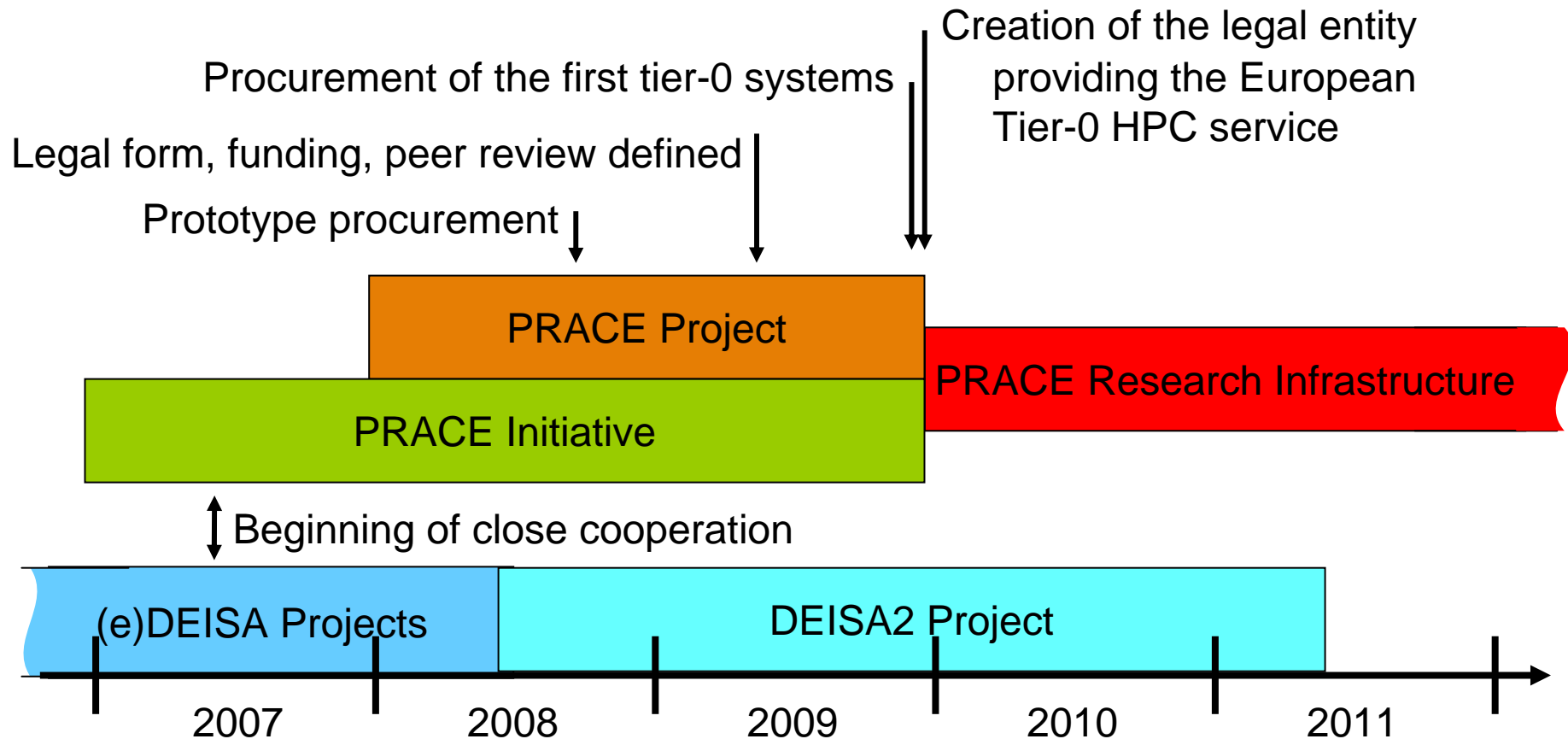
Provide training and education and disseminate results



The next tasks (II/II): ... growing into a persistent Research Infrastructure

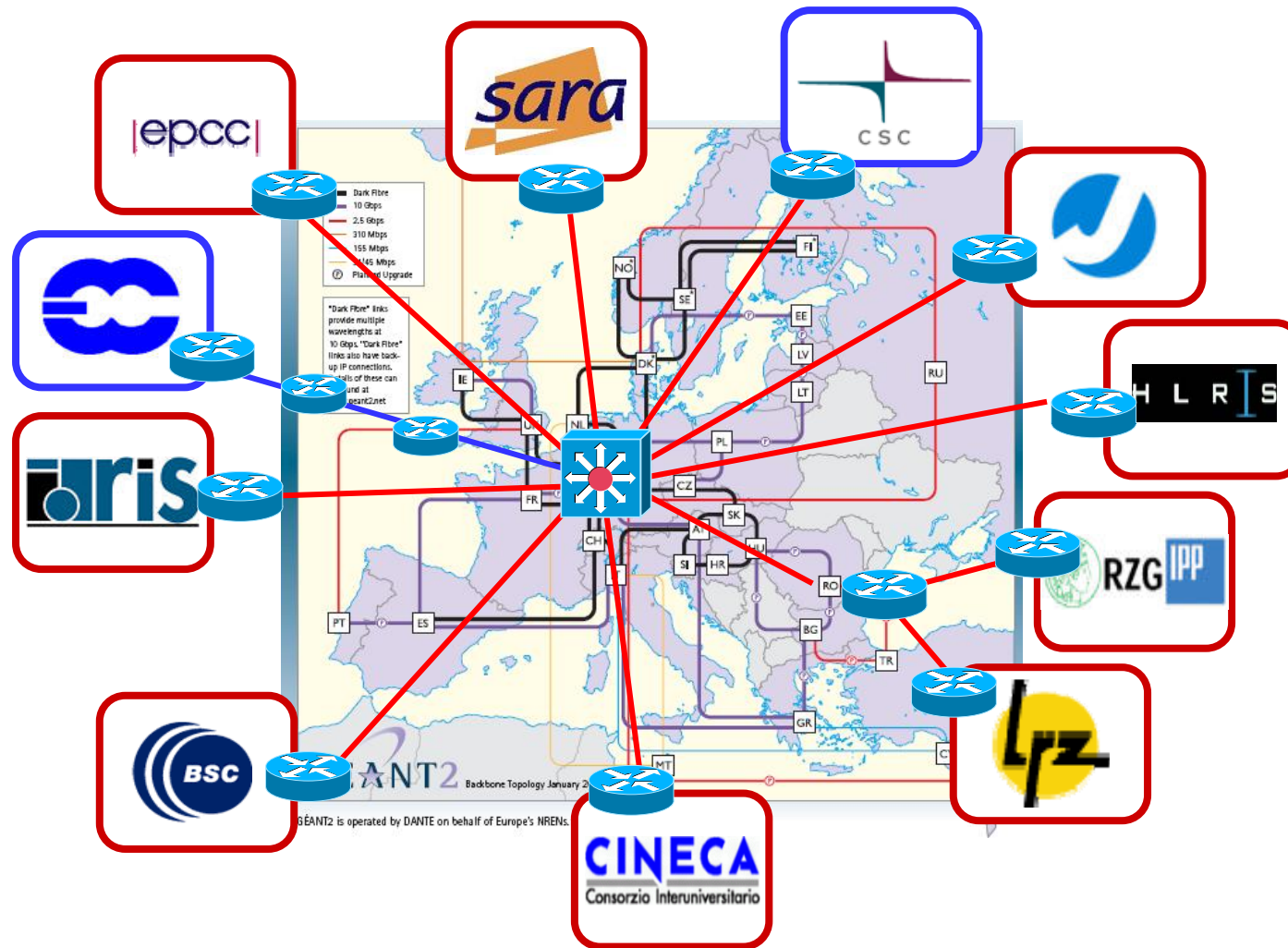
- Identify architectures and vendors capable of delivering Petaflops systems by 2009/2010 and install prototypes at partner sites to verify viability
- Define consistent operation models and evaluate management software
- Capture application requirements and create a benchmark suite
- Port, optimize and scale selected applications
- Define an open, permanent procurement process
- Define and implement a strategy for continuous HPC technology evaluation and system evolution within the RI
- Foster development of components for future multi-petascale systems in cooperation with European and international HPC industry
- Start process of continuous development and cyclic procurement of technology, software & systems for PRACE Research Infrastructure

DEISA and PRACE Roadmaps



- Goals of DEISA and PRACE are in-line with ESFRI roadmap
- DEISA and PRACE work together building up the European HPC ecosystem

Current DEISA network setup



- DFN
- FUNET
- GARR
- Rediris
- RENATER
- SURFnet
- UKERNA

Dedicated
10 Gb/s
wavelength

1 Gb/s LSP
GRE-Tunnel

Some application throughput measurements

Iperf program (TCP 10 s site-to-site)

- up to 4 Gb/s
- Dependent on system network options, cpu usage, system configuration, network layout (local and remote)
- Packet reordering can slow down transmissions dependent on used hardware equipment

GPFS (local throughput) → 16 Gb/s without any problems

GPFS (remote transfer) →

Teragrid test (2004) →

**30 Gb/s pipe filled using 40 servers and 64 clients
(highly dependent on system and network layout)**

HPC application demands on networking

Simple demands are:

A transparent multiple network domains spanning, high throughput, error free, low latency data pipe!

- **Most users don't care about network transport protocols. They use the network.**
- **But default applications often don't get anticipated throughput, though network is !! free !!**
- **There are application programmers, which are aware of the network, but they need support. E.g. „no“ command at AIX has about 132 options. Which one to use and how to set and what if my partner has any other OS?**

Future infrastructure needs for HPC computing

- **Operation and monitoring of network links across several administrative domains is a challenge**
- **GEANT2 and the NRENs have done a good job here providing the infrastructure for DEISA ... at least so far**
- **But things are becoming more complicated in the future having virtual organizations building up and being suspended in even shorter time frames**
- **Providing network services for upcoming and always changing grid infrastructures will become a new challenge. There are also those AAA issues.**
- **Having a secure and dedicated infrastructure like DEISA allows to rest easy**

But what and how to improve in future?

Future network requirements for HPC

- **Optical protection of links**
- **Bandwidth on Demand (Adhoc and in Advance) services**
- **Cross Domain link management including AAA**
- **Automatic management of several OSI layers**
- **Automatic reconfiguration for optical links working in local / remote environments (using OADMs)**
- **...**

I know much of this is already on the way, but there is a lot more to do

Summary and Conclusion

Since DEISA started in 2004 and the PRACE project started in 2008,

- a lot of work has been done in network research, as well in hardware as also in software development,
- new network management tools, network protocols, reservation systems, and applications using the underlying network have been developed.
- But fully utilizing the already available network resources with currently existing HPC applications without deep network expertise of application programmers could not be achieved.
- Optimal network usage is furthermore a challenge.
- Things have been improved, e.g. hybrid networks come up more and more, but there is a lot of work out there, that has to be done.

Questions ???

