SEVENTH FRAMEWORK PROGRAMME
Research Infrastructures

INFRA-2011-2.3.5 – Second Implementation Phase of the European High Performance Computing (HPC) service PRACE

PRACE-2IP

PRACE Second Implementation Project

Grant Agreement Number: RI-283493

D10.1
First Annual Report of WP10

Final

Version: 1.0
Author(s): Andreas Schott, GCS/RZG
Date: 27.08.2012

Project and Deliverable Information Sheet

PRACE Project	**Project Ref. №: RI-283493**	
	Project Title: PRACE Second Implementation Project	
	Project Web Site: http://www.prace-project.eu	
	Deliverable ID: < **D10.1**>	
	Deliverable Nature: <DOC_TYPE: Report / Other>	
	Deliverable Level: PU*	**Contractual Date of Delivery:** 31/08/2012
		Actual Date of Delivery: 31/08/2012
	EC Project Officer: ***Leonardo Flores Añover***	

* - The dissemination level are indicated as follows: **PU** – Public, **PP** – Restricted to other participants (including the Commission Services), **RE** – Restricted to a group specified by the consortium (including the Commission Services). **CO** – Confidential, only for members of the consortium (including the Commission Services).

Document Control Sheet

Document	**Title: First Annual Report of WP10**	
	ID: D10.1	
	Version: <1.0 >	**Status: *Final***
	Available at: http://www.prace-project.eu	
	Software Tool: Microsoft Word 2007	
	File(s): D10.1.docx	
Authorship	**Written by:**	Andreas Schott, GCS/RZG
	Contributors:	Jules Wolfrat (SARA), Luigi Calori (CINECA), Gabriele Carteni (BSC), Tilo Eißler (LRZ), Ender Güler (UHeM), Giuseppe Fiameni (CINECA), Ilya Saverchenko (LRZ)
	Reviewed by:	St. Janetzko, FZJ; D. Erwin, PMO
	Approved by:	MB/TB

Document Status Sheet

Version	Date	Status	Comments
0.1	27/June/2012	Draft	TOC, Responsible Persons
0.2	11/July/2012	Draft	Initial Texts
0.3	27/July/2012	Draft	Contribution Jules
0.4	30/July/2012	Draft	Contribution Ender
0.5	31/July/2012	Draft	Contribution Giuseppe
0.6	01/August/2012	Draft	Contribution Tilo, Gabriele
0.7	06/August/2012	Draft	Contribution Ilya, Luigi
0.8	09/August/2012	Draft	Changes proposed by Jules, Acronyms by Ender added

0.9	13/August/2012	Draft	Mainly formatting and many formulation improvements
0.10	18/August/2012	Updated Draft	Updated contribution from Luigi
0.11	21/August/2012	Updated Draft	Updated contribution from Jules
0.12	22/August/2012	Updated Draft	Modifications for T10.3 by Markus
0.13	23/August/2012	Updated Draft	Additions for HPSS (testing results) and iRODS (description and workshop agenda update)
0.14	25/August/2012	Updated Draft	Accepted changes and modified throughout the whole deliverable according to Dietmar's comments
0.15	27/August/2012	Updated Draft	Executive summary modified further
0.16	27/August/2012	Finalized Draft	Prepared for formatting, removing deleted text, dropping obsoleted comments, leaving those which should help in judging the modifications
1.0	27/August/2012	Final version	Cleanup for printable version

Document Keywords

Keywords:	PRACE, HPC, Research Infrastructure

Table of Contents

Project and Deliverable Information Sheet i
Document Control Sheet i
Document Status Sheet i
Document Keywords iii
Table of Contents iv
List of Figures v
List of Tables v
References and Applicable Documents v
List of Acronyms and Abbreviations vi
Executive Summary 1
1 Introduction 2
2 Enhancing the Existing Infrastructure 3
2.1 Authentication and Authorization 3
2.2 Accounting 3
2.2.1 *Central accounting repository* 3
2.2.2 *Storage Accounting and Reporting* 4
2.3 DECI portal 4
2.4 PRACE-Service-Portal 5
2.5 Inca-Monitoring improvements 6
2.6 Collaboration with other technological oriented activities 6
3 Evaluating Data-Services 8
3.1 PRACE Data Strategy 8
3.2 File-Transfer-Technologies 9
3.3 iRODS – integrated Rule Oriented Data System 10
3.4 File System Technologies 12
3.4.1 *Disk-Oriented File-System-Technologies by PANASAS* 12
3.4.2 *Hierarchical Storage Management: Automatic Archiving with HPSS and GHI* 13
4 Remote Visualization 16
4.1 Introduction 16
4.2 State-of-the-Art 18
4.2.1 *Scientific visualization applications supporting Client-Server paradigm* 18
4.2.2 *Application neutral, session oriented, VNC-like solutions* 19
4.3 Existing Partner Visualization Services 20
4.4 CINECA pilot project: GUI manager for a remote visualization TurboVNC session using PBX job scheduler. 20
4.5 *CINECA Summer school of Scientific Visualization* 21
5 Annex 22
5.1 Storage Accounting Questionnaire 22
5.2 Comparison of HPC-Europe and CINES tools for proposal management 23
5.3 iRODS-Workshop Preliminary Agenda 27
5.4 Functionality, Performance and Failover/Recovery of HPSS and GHI 29
5.4.1 *HPSS Functionality Tests* 29
5.4.2 *HPSS Performance Tests* 29

5.4.3 *HPSS Failover and Recovery Tests* *30*
5.4.4 *GHI Functionality Tests* *31*
5.4.5 *GHI Performance Tests* *31*
5.4.6 *GHI Failover and Recovery Tests* *31*
5.5 Remote Visualization Pilot Project at CINECA 32
5.5.1 *Requirements* *32*
5.5.2 *Allocated resources and deployment constraints* *32*
5.5.3 *Remote visualization layer* *32*
5.5.4 *Deployment setup* *33*
5.5.5 *Remote Connection Manager* *33*
5.5.6 *Deployment on PLX cluster* *35*
5.5.7 *Evaluation and further development* *35*

List of Figures

Figure 1: Schematic setup for HPSS 14
Figure 2: Detail configuration and data flow with HPSS 14
Figure 3: GHI: GPFS interface to HPSS 15
Figure 4: Generic configuration of a remote visualization system setup 16
Figure 5: Components of a generic browser-based application setup 17
Figure 6: Video-streaming configuration 17
Figure 7: Video-streaming configuration 18
Figure 8: Schematic login process for starting remote visualization 33
Figure 9: Login via Remote Connection Manager 33
Figure 10: Main Panel for Remote Connection Manager 34

List of Tables

Table 1: File Data Transfers use cases. 9
Table 2: Remote Visualization technologies in use at the different PRACE sites 19
Table 3: Existing partners' visualization services 20

References and Applicable Documents

[1] http://www.prace-project.eu
[2] EMI project: http://www.eu-emi.eu/home
[3] eduPerson schema: http://middleware.internet2.edu/eduperson/docs/internet2-mace-dir-eduperson-201203.html
[4] SCHAC schema: http://datatracker.ietf.org/doc/rfc6338/
[5] GridSAFE: http://gridsafe.forge.nesc.ac.uk/Documentation/GridSafeDocumentation/
[6] HPC-Europa2 project: http://www.hpc-europa.org/
[7] IGE project: http://www.ige-project.eu/
[8] European Middleware Initiative - Storage Accounting Record Proposal: http://www.ggf.org/Public_Comment_Docs/Documents/2012-02/EMI-StAR-OGF-info-doc-v2.pdf
[9] Globus Online: http://www.globusonline.org
[10] Unicore FTP: http://www.unicore.eu/documentation/manuals/unicore6/files/uftp-1.1.0/uftp-manual.html
[11] BBCP, http://www.slac.stanford.edu/~abh/bbcp/
[12] EGI (European Grid Infrastructure): http://www.egi.eu

[13] MAPPER: http://www.mapper-project.eu
[14] Globus Toolkit: http://www.globus.org/toolkit/
[15] MAPPER project deliverables: http://www.mapper-project.eu/web/guest/documents
[16] EMI: http://www.eu-emi.eu/
[17] PRACE iRODS workshop: https://www.irods.org/index.php/Introduction_to_iRODS
[18] Agenda iRODs-workshop: http://www.prace-project.eu/iRODS-workshop
[19] VirtualGL presentation: http://www.cineca.it/sites/default/files/VirtualGL.pdf
[20] VirtualGL: http://www.virtualgl.org/
[21] Remote Visualization LRZ: http://www.lrz.de/services/compute/visualisation/
[22] Remote Visualization RZG: http://www.rzg.mpg.de/visualisation/remote-visualization
[23] Remote Visualization SARA: http://www.sara.nl/systems/visualization/usage
[24] IBM-Whitepaper: https://www-304.ibm.com/partnerworld/wps/servlet/ContentHandler/whitepaper/systemx/linux_windows/install/lc=en_US
[25] ThinAnywhere: http://www.thinanywhere.com/
[26] NICE-DCV software: http://www.nice-software.com/products/dcv
[27] PCoIP: http://www.teradici.com/pcoip/pcoip-technology.php
[28] CINECA Remote Visualization Summer School: http://www.cineca.it/page/summer-school-scientific-visualization
[29] CINECA Remote Visualization Summer School agenda and presentation material: http://www.cineca.it/page/agenda-and-presentation-material
[30] SVN repository of Remote Connection Manager: https://hpc-forge.cineca.it/svn/RemoteGraph/trunk/
[31] Remote Connection Manager User Documentation: http://www.hpc.cineca.it/content/remote-visualization
[32] Inca home page: http://inca.sdsc.edu/drupal/
[33] EUDAT-project: http://www.eudat.eu
[34] VisIT: http://wci.llnl.gov/codes/visit
[35] ParaView: http://www.paraview.org
[36] IDL: http://www.excelisvis.com/idl
[37] StarCCM: http://www.cd-adcapo.com/products/star_ccm_plus

List of Acronyms and Abbreviations

AAA	Authorization, Authentication, Accounting.
AAI	Authentication and Authorisation Infrastructure
API	Application Programming Interface
BAdW	Bayerischen Akademie der Wissenschaften (Germany)
BSC	Barcelona Supercomputing Center (Spain)
CINECA	Consorzio Interuniversitario, the largest Italian computing centre (Italy)
CINES	Centre Informatique National de l'Enseignement Supérieur (represented in PRACE by GENCI, France)
CPU	Central Processing Unit
DECI	Distributed European Computing Initiative
DEISA	Distributed European Infrastructure for Supercomputing Applications. EU project by leading national HPC centres.
DMOS	Distributed Maintenance Organisation System, tools for managing maintenance information
DPMDB	DECI Project Management Database
EGI	European Grid Infrastructure
EMI	European Middleware Initiative

EPCC	Edinburg Parallel Computing Centre (represented in PRACE by EPSRC, United Kingdom)
FC	Fiber Channel
FZJ	Forschungszentrum Jülich (Germany)
Gb/s	Giga (= 10^9) bits per second, also Gbit/s
GB/s	Giga (= 10^9) Bytes per second
GCS	Gauss Centre for Supercomputing (Germany)
GDDR	Graphic Double Data Rate memory
GENCI	Grand Equipement National de Calcul Intensif (France)
GPFS	General Parallel File System, a high performance file-system by IBM
GHI	GPFS to HPSS interface, providing HSM functionality to GPFS
GPU	Graphic Processing Unit
GUI	Graphical User Interface
HBA	Host Bus Adapter
HCA	Host Channel Adapter
HPC	High Performance Computing; Computing at a high performance level at any given time; often used synonym with Supercomputing
HPSS	High Performance Storage System, highly scalable tape management system for handling huge amount of data
HSM	Hierarchical Storage Management, transparently moving data between disk and tape storage devices
IB	InfiniBand
IBM	Formerly known as International Business Machines
IDRIS	Institut du Développement et des Ressources en Informatique Scientifique (represented in PRACE by GENCI, France)
IGE	Initiative for Globus in Europe
I/O	Input/Output
iRODS	integrated Rule Oriented Data System
ISTP	Internal Specific Targeted Project
JSC	Jülich Supercomputing Centre (FZJ, Germany)
KTH	Kungliga Tekniska Högskolan (represented in PRACE by SNIC, Sweden)
LANL	Los Alamos National Laboratory, Los Alamos, New Mexico (USA)
LBNL	Lawrence Berkeley National Laboratory, Berkeley, California (USA)
LDAP	Lightweight Directory Access Protocol
LLNL	Laurence Livermore National Laboratory, Livermore, California (USA)
LRZ	Leibniz Supercomputing Centre (Garching, Germany)
LUN	Logical Unit Number for addressing storage
MAPPER	Multiscale Applications on European e-Infrastructures; FP7-Project
MoU	Memorandum of Understanding
NAS	Network-Attached Storage
NCF	Netherlands Computing Facilities (Netherlands)
NFS	Network File System
OpenGL	Open Graphic Library
ORNL	Oak Ridge National Laboratory, Oak Ridge, Tennessee (USA)
OS	Operating System
PB	Peta (= 10^{15}) Bytes (= 8 bits), also PByte
PBS	Portable Batch System, a widely used batch scheduler
PCoIP	PC over Internet
pNFS	Parallel Network File System
POSIX	Portable OS Interface for Unix

PPR	PRACE Peer Review DB
PRACE	Partnership for Advanced Computing in Europe; Project Acronym
PSNC	Poznan Supercomputing and Networking Centre (Poland)
QDR	Quad Data Rate
RZG	Rechenzentrum Garching (RZG) of the Max Planck Society and the IPP (Max Planck Institute for Plasmaphsics), Germany
SAN	Storage Area Network
SARA	Stichting Academisch Rekencentrum Amsterdam (Netherlands)
SAS	Serial Attached SCSI
SATA	Serial Advanced Technology Attachment (bus)
SLA	Service Level Agreement
SNIC	Swedish National Infrastructure for Computing (Sweden)
SNL	Sandia National Laboratories, Albuquerqe, New Mexico (USA) and Livermore, California (USA)
StAR	Storage Accounting Record
STS	Security Token Service
Tier-0	Denotes the apex of a conceptual pyramid of HPC systems. In this context the Supercomputing Research Infrastructure would host the Tier-0 systems; national or topical HPC centres would constitute Tier-1
UHeM	National Center for HPC of Turkey (Formerly UYBHM)
UNICORE	Uniform Interface to Computing Resources; grid software for seamless access to distributed resources
VNC	Virtual Network Computing. A graphical desktop sharing system that transmits the keyboard and mouse events from one computer to another, relaying the graphical screen updates back in the other direction, over a network

Executive Summary

The major aim of Work Package 10 (WP10) 'Advancing the Operational Infrastructure' is promoting the software technology and services needed for the operation of the integrated infrastructure inherited from the DEISA project and the PRACE preparatory phase. After finalization of the PRACE-1IP project, WP10 is also continuing the work of the technology task T6.3 of the work package WP6 of PRACE-1IP. This activity is split into three tasks which cover different aspects of the goal to be achieved. The first one is focused on the existing infrastructure. The second one concentrates on data services, which go beyond the already existing ones, because access to data is an essential need for all scientists doing computations within PRACE. The third task deals with the remote visualization of data. This can help in reducing the amount of data to be transferred and also enable users to adapt parameters for future computations according to the outcome of the visualization.

The first task, enhancing the existing infrastructure, is aligned with the different services provided, while currently the main focus is related to authentication and authorization, accounting, monitoring and interfaces to the infrastructure. In order to stay aligned with other important projects in these fields, a sub-task is taking care of collaborations with other technological oriented projects. Some of the activities started only after the repective work of T6.3 in WP6 of PRACE-1IP has ended and will now be continued and completed in WP10 of PRACE-2IP. The unification of the users' view on the Tier-0 and the Tier-1 infrastructure, especially for the submission of proposals and their different review processes is a main goal in the DECI-portal activity part of this task. The requirements figured out by WP2, caring for the DECI calls, the needs of AISBL responsible for the Tier-0 calls, and the wish for a common tool for both, Tier-0 and Tier-1 calls, implied an intensive investigation in the possible solutions. It is planned for the PRACE all-hands-meeting in Paris in September to achieve a decision based on technical and non-technical arguments, which tool to select on which then the adjustment and development work will start.

For the second task, data services, it turned out, that there is a need for a strategic planning, on how to deal with data in PRACE in general. The current model assumes all input and output data related to computations performed in Tier-0, but also in most cases in Tier-1, to be transferred into and out of the PRACE systems in a relatively short time frame. Considering the ever increasing amount of such data and the limited external network capabilities this way of data handling encounters more and more problems, although improvements in the file transfer technolgies are achieved. Thus the management agreed that WP10 will collect options and summarize them as possible methods for the treatment of data in PRACE. Beside these general considerations the evolution of file transfer technologies are continuosly followed. Reflecting already existing users' requests in the latest DECI-calls, iRODS was selected as the repository-technology to investigate. Concerning the evaluation of file systems the main focus was set on hierarchical storage management for long-term storage needs.

The third task, remote visualization, has made progress in collecting information about state-of-the-art remote-visualization technologies and surveying their deployment at PRACE partner sites. Selected technologies, as well as specific software tools for the visualization and analysis of scientific data were presented to an audience of scientific users at a summer school. Furthermore, a pilot project, which addresses some of the shortcomings of existing, out-of-the-box remote visualization infrastructures has been implemented at one partner site. The project aims at improving the scalability of the visualization infrastructure and in particular the ease-of-use for scientific users end-users concerning client setup and connection handling. It is being prepared for export to other sites collaborating within this task for their further assessment concerning suitability for the corresponding production environments.

1 Introduction

The objectives of WP10 are:

- Enhancing the existing Tier-1 operational infrastructure
- Evaluation of additional data services
- Remote Visualization

Each of these objectives has a corresponding task in the work-package. Where appropriate the tasks are subdivided in sub-tasks to better address the specific activity.

Structure of the Document

The following document consists of three more chapters, one for each of the tasks addressing one of the objectives listed above. The single chapters then contain several sections covering the work of the respective sub-tasks, which are logically mainly independent from each other. Finally in an appendix-chapter several sections provide even more detailed or additional information for some of the tasks or sub-tasks.

Relation to WP6 Operations in PRACE

WP6 is responsible for the operation of the infrastructure of and the services provided in PRACE. In PRACE-1IP the technological evolution was covered in a task part of WP6, while in PRACE-2IP a separate work-package WP10 is dealing with new technological developments. But it is extremely important to correlate the work of the two work-packages, so that requirements showing up in WP6 can be adequately be addressed in WP10. For this reason the all-hands meetings of WP6 and WP10 are arranged jointly. There have been two of such all-hands meetings, one in October 2011 in Amsterdam and the other in April 2012 in Bologna. In these meetings the direct requests of WP6 to WP10 as well as those of users have been addressed and the different sub-tasks have planned for their respective activities. Both meetings had around 40 participants.

2 Enhancing the Existing Infrastructure

The objective of task 10.1 is to evaluate the existing services as inherited from DEISA and identify and evaluate options for technical enhancements, also using the results of user surveys. This includes the continuation of the work done by the WP6.3 task in PRACE-1IP.

2.1 Authentication and Authorization

This activity addresses possible enhancements in Authentication and Authorisation Infrastructure (AAI) technologies to access the Tier-1 and Tier-0 infrastructure.

The exchange of personal data for authorisation is an important topic for service providers working with user communities. These communities will be able to form an Authentication and Authorisation Infrastructure (AAI) federation, where personal data is managed in a structured way using internally defined attribute schemas or standardized attribute schemas, like the eduPerson schema [3] and the SCHAC schema [4]. These federations can be organized on a national on an international level. The use of already existing information can improve both the registration procedure of users of the PRACE infrastructure as well as the ease of access to the resources. To enable this use of attribute information provided by AAI federations several issues have to be resolved. First of all, the attribute information must fit the requirements of the service provider. For PRACE additional information is needed, like to which resources the user is granted access. Such information has to be maintained by PRACE. For PRACE a merge of external and internal attribute information will be needed if external information is used. Currently PRACE maintains all attribute information about users in the PRACE LDAP based user administration repository. In addition to the technical problem of using external attribute data there is also the issue of trust between AAI federations and service providers like PRACE. The service provider must have trust in the way attribute information is validated and maintained and the attribute provider must have trust that the service provider will behave in a secure way with the provided data, e.g. protecting the privacy of the user.

There are several initiatives to improve the exchange of attribute data between AAI federations and service providers. For instance the European E-infrastructure Forum (EEF) organized several meetings to discuss the collaboration between AAI federations and service providers. And GÉANT produced a draft Code of Conduct for service providers, which if accepting this Code can use attribute data from AAI federations. PRACE follows the developments in this area by visiting meetings organized by EEF and GÉANT on these issues. There are currently no activities planned for the evaluation of using external attribute information through pilot projects or the like.

In T6.3 of PRACE-1IP the developments for the Security Token Service (STS), as proposed by the EMI project [2], were evaluated. This will be evaluated further once available and if suitable use cases are identified.

2.2 Accounting

2.2.1 *Central accounting repository*

Task WP6.3 of PRACE-1IP has successfully completed the evaluation of the set-up of a central repository for the provision of accounting information to users, Principal Investigators and site administrators of the PRACE infrastructure. This repository is based on the Grid-SAFE tools [5] as developed and maintained by the PRACE partner EPCC. The last step is

that an ISTP document must be produced as input for the final acceptance as a production facility. Task 10.1 will produce the ISTP.

2.2.2 *Storage Accounting and Reporting*

The objective of this activity is to enhance the PRACE resource accounting infrastructure by analyzing the need of storage accounting and disk usage information for users.

A survey has been prepared to collect information from all PRACE partners and AISBL on this subject. There are 27 questions in total. The survey has the following sub-titles:

- Current status of partner sites: Questions in this part are related to current status of sites on this subject. Relations between disk usage and users/projects are questioned. An attempt was made to get information about the current disk usage reporting and/or storage accounting tools at each partner site.
- Policies and requirements: Disk usage reporting and storage accounting requirements and policies were discussed in this section of survey.
- Implementation related questions: To steer the studies on storage accounting in PRACE, the objective is to get as much information as possible. In relation to that, implementing a storage accounting infrastructure and storage accounting records are discussed in non-detailed fashion around the European Middleware Initiatives Storage Accounting Record Proposal (StAR) [8].

All partners/sites are asked to fill in the survey document until end of October 2012. The results will go into a requirements document until the beginning of the next year. The most prominent requirements identified within this document will then be used to eventually define further actions:

- Provide requirements to EMI and possible other middleware development activities by other initiatives;
- If interesting solution exist these will be tested at partners;
- If needed, adapt existing solutions for use in PRACE.

2.3 DECI portal

For the management of the DECI proposals and projects the DPMDB facility, developed by DEISA, is used. WP2, responsible for the management of the DECI calls within PRACE-2IP, has asked WP10 for some enhancements to the facility. The management of the proposals from submission to final acceptance is labour intensive and error prone. In the joint PRACE-2IP-WP6/WP10 all-hands meeting in Bologna in April 2012 WP2 presented the list of major problems with the current tool:

- Time consuming and not scalable especially when the frequency of calls is increased to two per year;
- Error prone, due to continuous copy and pastes needed;
- Continuous adaptation of different forms. Change in one form implies changes in other forms;
- Communication through e-mail is cumbersome (extracting addresses from proposals is needed);
- Documents are difficult to manage (naming conventions, storage locations in BSCW);

WP2 also presented a list of necessary or desirable improvements:

- Electronic submission of proposals via a Web Portal;
- Technical Evaluation (TE) via Web Portal, where all relevant data from proposal are visible in the TE form, which is used by DECI staff to provide feedback on the technical feasibility of the proposals;
- Scientific Evaluation (SE) via Web Portal where evaluators can get limited access to relevant proposals and TE.
- Documents are stored in an database integrated in the Web Portal ;
- Management of access rights to information;
- Flexibility in redesigning Web Portal and integrated DB;
- User friendly GUI to the Web Portal;
- Connection between Web Portal DB and DPMDB or similar solutions where the proposal databases is connected to the database which keeps track of the progress of projects;
- (Part of) communication with Applicants via Web Portal;
- More flexibility in DPMDB;
- Redesign/Restructuring of DPMDB;

WP2 also explained on how they envisioned the required and desirable interactions with the new portal and the databases behind it. The expected advantages of the proposed set-up are:

- More efficient and faster way of working;
- Get rid of a lot of paperwork;
- Reduce copy – paste actions substantially, making it less error prone;
- Better consistency of data (applicant addresses and affiliations);
- Give applicants one single point of access to all data/ information;
- Easier and more efficient to perform TE and SE;

Two existing tools were proposed as possible candidates for the desired functionalities:

- HPC Europa tool including DB – This tool is used by the HPC-Europa2 project [6] and is supported by the PRACE partner CINECA;
- Tier-0 peer review tool and DB (PPR) – This tool is used for the management of Tier-0 proposals and is supported by the PRACE partner CINES.

For both tools it must be investigated if they can be adapted to the requirements of WP2 and what support is available. As a first step for both tools dedicated meetings by video conference were scheduled with WP10 members where CINES and CINECA presented the features of the respective tools and what possibilities there are for adaptations. A comparison of the two tools has been made, see annex 5.2. This comparison will be used to take a technical decision at latest at the all-hands meeting in Paris in September 2012, including WP2 and people from the Tier-1 non-technical arguments into the decision process.

There should be one tool to manage all PRACE proposals, both for Tier-0 and Tier-1 (DECI) access. The PPR tool is already used by PRACE for Tier-0, so the technical preference currently is to extend this facility also for use by the DECI calls.

2.4 PRACE-Service-Portal

PRACE users require a wide variety of information in order to efficiently utilize the services offered in the e-Infrastructure. This information includes network status and performance, HPC resource maintenance schedule, Grid service availability and functionality and so on. Over the course of the PRACE-1IP project shortcomings in information provisioning and distribution were identified, as most of the collected information is available solely to PRACE

staff members. Annual PRACE user surveys emphasise these limitations and provide details on the kind of data users need to efficiently work in the PRACE e-Infrastructure.

PRACE-2IP WP10 defined a task to evaluate, design and implement a portal for providing users with information on the PRACE e-Infrastructure, such as availability, accessibility and performance of PRACE resources and services. The main goal of this task is to address user requirements by providing desired functionality based on existing or novel technologies.

Work performed by this task is mostly based on requirements collected through multiple user surveys performed by PRACE, such as those by PRACE-1IP-WP6.3. As a result the items to be provided with the highest priority are:

- Common Production Environment status. The information is available through Inca, but not publicly published;
- Queue waiting time.
- Resource maintenance information. Planned to be provided through the DMOS facility, which was developed under DEISA;
- Accounting information. Discussed in section 2.2.

2.5 Inca-Monitoring improvements

Within PRACE-1IP WP6 and PRACE-2IP WP6, the user-level monitoring application Inca [32] has been setup to monitor the PRACE infrastructure services. It has been successfully utilized before within the DEISA project. In DEISA the Inca application setup included an authentication mechanism to limit access to sensitive data, such as user names and internal addresses of resources and services. The authentication mechanism was based on a manually managed access control list.

Utilizing an access list has been sufficient within DEISA. It is a solution capable of dealing well with a small amount of users. As the number of users and information significantly increased within PRACE, the manual administration of the access list got unfeasible. Furthermore, checking the access list causes a high load on the monitoring server.

As part of this sub-task, a solution to overcome these limitations has been developed. It is based on the fact that the user data required for authentication is already available in the PRACE LDAP. To leverage this, an interface for connecting the Inca authentication to the LDAP has been developed. This solution avoids the management of a separate source for authentication information. It is currently in internal testing stage and will be transferred into production soon.

2.6 Collaboration with other technological oriented activities

Fostering the collaboration with other EU projects and e-Infrastructure is an important activity which aims for better supporting user communities and strengthening the collaboration with external technology providers. The collaborations with external organisations have been initiated in the course of PRACE-1IP and are continued within this work package since July 2012. Since these collaborations are mainly technologically oriented WP10 is maintaining them reflecting users' needs reported especially by WP6.

Current collaborations include:

- **MAPPER**: The MAPPER project (Multiscale APPlications on EuRopean e-infrastructures) [13] aims at deploying a computational science environment for distributed multi-scale computing, on and across European e-Infrastructures, including PRACE and EGI. The collaboration between the two projects initiated in May 2011 and

was coordinated via a Task Force comprising specialists from each of the three organisations (MAPPER, PRACE, EGI-Inspire). The task force is now working to integrate PRACE and EGI user support services in order to provide end users with a centralized interface for submitting support requests.

- **EMI**: The EMI (European Middleware Initiative) [16] project is a close collaboration of the four major European middleware providers, ARC, dCache, gLite and UNICORE. Its aim is to deliver a consolidated set of middleware components for deployment in EGI, PRACE and other DCIs, extend the interoperability and integration between grids and other computing infrastructure. The collaboration with the EMI project initiated on September 2011 to define a common framework of collaboration where to exchange expertise, support the evolution of UNICORE components, access to emerging technologies, enforce the sustainability of adopted technology. A joint work-plan to implement collaboration's objectives was defined in a Memorandum of Understanding (MoU) which is currently under discussion within respective coordination bodies. The PRACE AISBL will sign the MoU if a consensus on defined objectives is reached. Features covered in the MoU are:
    - Support and new requirements for the UNICORE tools;
    - Evaluation of the Security Token Service (STS);
    - Development of an SLA for operational support;
    - Support for training and dissemination.
- **IGE:** The Initiative for Globus in Europe (IGE) [13] is a project supporting the European computing infrastructures by providing a central point of contact in Europe for the development, customisation, provisioning, support, and maintenance of components of the Globus Toolkit [14], including GridFTP and GSI-SSH which are currently deployed in PRACE. As for EMI, a joint work-plan was defined within a MoU currently under discussion within respective coordination bodies. Planned features of the MoU are:
    - Evaluation of Globus Online and GridWay services;
    - Support for Globus tools in use by PRACE;
    - Support for training and dissemination.
- EGI: The collaboration with the European Grid Infrastructure (EGI) [12] has been intensified around the MAPPER project since MAPPER's objectives are focused on implementing a number of significant use cases for collaboration and interoperation between PRACE and EGI. These use cases require the joint use of PRACE and EGI computational resources. Currently we are working on the exchange of resource usage information between EGI and PRACE and interoperability of the helpdesks of the infrastructures.A meeting among infrastructure experts will be held during the EGI Technical Forum 2012 (Prague, Czech Republic, September 17-21).

3 Evaluating Data-Services

All HPC systems need fast data storage for the data used in the calculations to be performed. This covers input and output data, and locally most important scratch or restart data which is not required once the calculation is completed. This temporary data is stored on parallel file systems which are usually provided together with the HPC system by the vendor. Thus there is in general not much of a choice.

For non-temporary data, output data exceeds input data in size in most cases by much and such data usually also needs to be stored for a longer time, either to continue calculation later or to perform post-processing. Here several different solutions are possible. One can consider storing the data on an independent file system, e.g. if the post processing is also performed on the site where the data has been produced. This scenario is then also a motivation for remote visualization. Furthermore, long-term storage is an important feature which is realized by file systems using hierarachical storage methods, where data is transparently moved out of the disk storage to tape devices and back from tapes into the disk storage again. Section 3.4 deals with the evaluation of such an HSM file system.

There have been requests of DECI users for accessing data stored in iRODS, thus this technology has been identified as necessary to be supported by PRACE. This is addressed by the sub-task concerned with iRODS and the outcome is described in section 3.3.

It is also very important to transfer data into and out of PRACE as fast as possible, thus investigation in data transfer tools is followed up in section 3.2. This activity already started in PRACE-1IP and is continued here, while methods useful in Tier-1 as well as in Tier-0 are investigated.

All these different options of storing or transferring data to, from and withing PRACE, it is very important to also get a clear view, on how data should be treated in PRACE. For this reason a small group of experts will very openly consider about the possible option and provide the management level of PRACE with input for strategic decisions. This is outlined in more detail in section 3.1.

3.1 PRACE Data Strategy

The main focus of PRACE is the provision of high performance computing resources. But it clearly turns out that in general the more compute intensive calculations are performed the more resulting data is generated. In DEISA all users had their so called “home”-site, which was connected to the 10Gbit/s dedicated network infrastructure covering all Tier-1 sites. With PRACE the number of sites increased dramatically and an increasing amount of partners being the hosting site for the DECI users are connected only by usually much lower bandwidth connections. Thus it becomes more and more difficult to easily transfer huge amounts of data to a site, where the data can be stored for a longer period. This is even more dramatic for Tier-0, where all data has to be transferred out of the computing system in a short time frame.

At the moment users are given a relatively short time period, to transfer their output data from any of the PRACE computing systems to their personal or a global storage system. This seems to be very inefficient. Therefore it is important to find out better methods for the users to deal with their huge amount of data. Different options have to be evaluated technically, which range from the provision of long term storage inside PRACE to heavily improved network connections to all partners. Furthermore, there are also other EU-projects, as EUDAT [33], which investigate into possibilities for long term storage of data. Many sites involved in

EUDAT, also planning to provide long term storage, are connected to PRACE with high speed network connections. So many solutions, also based on collaborations, are thinkable; but all possible realizations touch strategic considerations in PRACE. Thus an agreement by the PRACE management was requested, to start consideration about possible strategies for the handling of data in PRACE.

Therefore, a small group of data-experts from five different PRACE-sites is collecting possible different options, and generate a review paper out of them. In this report the possible solutions will be outlined and weighted, possibly also considering costs. With help of this report the persons responsible for the future of PRACE shall be instrumented to take the most reasonable and viable decision on the strategy of the future handling of data in PRACE.

3.2 File-Transfer-Technologies

The objective of this subtask, part of the task 10.2 "Evaluating Data Services", is to investigate and evaluate alternatives or additional tools for the current high-performance file transfer service supported by PRACE and based on Globus GridFTP. This activity continues the work of PRACE-1IP, therefore it started only after PRACE-1IP has ended.

As consequence of the continuous growth in computing power, it is expected that many scientific data-intensive applications will produce tremendous amount of data and an even more challenging need for moving bulk data inbound and outbound of the PRACE Research Infrastructure.

High-performance wide-area bulk data transfer has always been a complex task for a variety of reasons, which are not only strictly related to the data size to be transferred. Improper configuration of the sending and receiving hosts, the capacity of the underlying storage and file systems, network congestion, in-path narrow links, software design issues, firewalls and local policies are all factors that affect the performance.

Concerns about bulk or big data transfers are confirmed and shared among scientific communities as well as DECI users. Moreover, significant efforts are being spent by other EU-funded projects, like EUDAT [33], for addressing new challenges raised in the big data management.

Table 1 shows the three main use cases for data transfer with PRACE sites involved, where the network connectivity plays an important role. But the behaviour of data transfer is not only affected by the underlying network, but also heavily dependend on the data topology (many small files, large single files, depth of directory trees, etc…).

The outcome and final results of a very similar activity carried out during PRACE-1IP (WP6-Task6.3) will be used to continue and extend this work by defining the methodology to be followed during the analysis as well as a list of tools to be tested.

Transfer Type	Source	Destination	Network
Inbound	External Site	PRACE Site	Public Internet
Outbound	PRACE Site	External Site	Public Internet
Internal	PRACE Site	PRACE Site	PRACE Network

Table 1: File Data Transfers use cases.

In order to evaluate different technologies, which are available today for scientific communities and provided for free, a methodology for performing technical tests has been defined. The methodology provides guidelines for setting up a test, rules for configuring a host (e.g. TCP tuning), policies to take into account network congestions that can occur during the tests and obviously a list of test cases to be executed and considering different datasets (many small files vs. few big files). The objective is to reproduce real use cases and compare all tools under the same operational conditions.

Reliability, community requirements, user interface, fault tolerance, code maturity and support will be also considered in addition to the sustainable throughput.

Tests are planned for the second year of PRACE-2IP and over the private network (transfers between PRACE sites) as well as the public network (transfers between a user host and a PRACE site and vice versa).

The list of tools will include the evaluation of Globus Online [9] which is currently under auditing by the PRACE Security Forum and comes with positive feedbacks from a preliminary evaluation carried out for Tier-0 systems within PRACE-1IP.

The pluggable file transfer mechanisms provided by UNICORE and recently extended with the Unicore FTP [10] or UFTP will be also taken into account and evaluated.

There is also an expression of interest for including BBCP [11] in the list of tools to be evaluated. BBCP offers different authentication mechanisms, including support for X.509 certificates, data parallelism and does not require a remote server running.

3.3 iRODS – integrated Rule Oriented Data System

The data management system iRODS aims at the transparent data and metadata handling with the goal of worldwide unified access to persistent data [17]. The users can access it through a unique identifier not knowing where the data actually is located. All such data handling is implemented by a Storage Resource Broker, which is able to interface different storage systems, disk oriented online space and tape oriented archives, and so can control the physical location of the data in an extremely flexible way.

Technological background on iRODS

The iRODS data management system provides transparent data and metadata handling across a system of distributed storage resources. In recent years the take-up of iRODS in large national and international projects has been very noticeable. iRODS is currently seen as a production ready data management system and is widely used in U.S, European and world wide projects including instance which are provided by national libraries, universities and organizations such as NASA. In addition to its evaluation within PRACE, iRODS is currently being investigated by other EU projects, most noticeably the FP7-funded EUDAT project which aims to provide a pan European data infrastructure.

The key features of iRODS, which we will discuss in more detail below, are scalability; flexibility, through the rule orientated design; integrated metadata; federation capabilities; and global data access via a wide range of clients.

Scalability: An iRODS instance is formed from a Metadata Catalog (iCAT), which stores state and descriptive information in a database, and a cluster of storage resources which may be distributed across several sites. This architecture allows for easy scaling, by simply adding an extra storage resource to the iRODS instance. What is more, iRODS can provide transparent access to existing file-systems such as GPFS via so called "mounted collections"

and can also be integrated with Mass storage systems such as HSM/HPSS to leverage the power of tape archives.

Rule Orientated Data Management: iRODS provides the ability for users, communities and administrators to define and realize their data management policies through sets of rules. These rules allow for the automation of administrative tasks, the enforcement of data management policies, and the implementation of workflows. The rules can be configured to run on a regular basis and/or be triggered by system events, such as data ingest, or the users themselves. This leads to a very powerful and flexible system where the business logic for users and communities alike can be embedded into iRODS thus providing a solution which is tailored to their needs.

Client Access: A wide range of client tools are available, which cover many different usage scenarios, and can be deployed on different systems. These provide the end users with a great deal of flexibility when it comes to using the iRODS system. A brief description of the major clients follows:

- icommands - Unix and Windows command line clients, simple and powerful set of tools for server and data management
- iDROP - swing based GUI desktop transfer manager, provides the ability to drag and drop files and also synchronize between desktop and iRODS servers
- Web Browser - Access iRODS servers through web browsers without the need to install client tools.
- iRODS Explorer for Windows - GUI browser
- WebDav - webdav enabled interface to iRODS via the DAVIS extension
- iRODS-FUSE - Mount iRODS collections locally for posix like access

Integrated Metadata: The metadata catalog (iCAT) provides state and user level metadata and can be used for administrative purposes, such as proofing data authenticity and producing audit trails, and user level activities alike. Individual users can assign metatdata attributes to their data and can use this metadata to selectively query their data. For example a user could tag data which originates from a certain sensor upon ingest and construct queries to return only data which came from that sensor.

Federation of Archives: The ability to easily federate multiple iRODS instances enables flexible architecture creation and simplifies collaboration since data from one iRODS instance can be opened to a user group which is support via another iRODS instance. This functionality ensures that archives can be linked without the need to move data or make large scale changes to the existing archive system. Moreover, the flexibility provided by the federation functionality allows archive architectures to be created such as master-slave systems or chained archives where data may be propagated through several linked archives.

The control over the flow and location of data is then achieved by the rules, which can be defined in iRODS. This configurable flexibility makes iRODS a very promising method for any data handling. Another FP7-funded project, EUDAT, which is data-oriented, therefore is also investigating deeply in iRODS.

PRACE Workshop on iRODS

Already some users doing calculations in DECI projects requested access to iRODS controlled data from within their calculations. It is apparent that these requests are not just driven by users who wish to learn how this could be achieved in principle, but also by users who would like to make use of data which resides in existing iRODS systems. Thus it is necessary not only to learn about, but also to deal with iRODS. For this reason a workshop

has been prepared, which will take place from Wednesday, 26th until Friday, 28th of September in Sweden at the Linköping University [18].

This open workshop will bring together people from a large range of disciplines interested in data management. The primary focus of this workshop will be the needs and requirements of the users in the HPC field and how they can be met using the iRODS technology. Participants will have the opportunity to gain a deeper understanding of the iRODS system, its features, and how it may be applied to solve their individual needs. The workshop will include an iRODS tutorial, demos, hands-on training sessions, presentations on user applications and will also tackle the iRODS strategy for the future. Attendees of the hands-on sessions will have the opportunity to install a fully functional iRODS data management system on their own laptop and can thus leave the tutorial with their own iRODS system.

3.4 File System Technologies

In all computer systems file systems are the basis for data storage. In computers with high parallelization such file systems have to provide such parallelisms also for the I/O-operations to the file systems. Since this is a very challenging requirement, vendors certify parallel file systems specifically for the HPC-system. Therefore there is usually not much of a choice for the file system attached to such a HPC-system. But there are a few parallel file systems, which can be attached in addition to the file system provided by vendor.

Such file systems may be useful for accessing data directly over the network, as it is provided by the MC-GPFS setup in DEISA inherited into PRACE for Tier-1 sites. Another aspect may be easy maintainability or automatic archiving to tape for long-term storage. It is clear that for such implementations PRACE needs to evolve its current policy that data will stay only for a short time period after a project has ended its calculations on a PRACE system. But as already explained above in section 3.1, due to the ever increasing amount of data produced and the limited external network capabilities, it may become necessary to reconsider this and eventually modify the PRACE strategy concerning data.

One specific file-system technology on the list of possible technologies to be evaluated in WP10 is Lustre. WP10 did not investigate herein, since WP12 already intensively worked on optimizing the performance of Lustre, which is reported in the deliverable D12.4 “Performance Optimized Lustre”.

For long term data storage file-systems, especially those with HSM functionality, are essential for transparently storing huge amount of data. Independently from the outcome of the strategy considerations within PRACE, evaluation of such file systems are useful for all computing centers having to deal with huge amount of data.

3.4.1 *Disk-Oriented File-System-Technologies by PANASAS*

In the very beginning contacts have been established to the vendor PANASAS, which is providing scalable disk-storage-solutions for HPC-systems, providing high parallelization for I/O-operations. This hardware would offer comfortable replacement functionality for defective disks as well as easy expandability for increased storage requirements.

The managing software is proprietary, so testing requires the provision of a test-system by the vendor. Negotiations had been started whether a reasonable agreement for a provision of such a test-system would be possible. In general provision of such a test-system was considered to be viable, but the possible location was difficult to match with the involvement of partners into this activity. Thus evaluating the PANASAS file system was postponed.

3.4.2 *Hierarchical Storage Management: Automatic Archiving with HPSS and GHI*

There are several technologies available and used in PRACE. Investigation in HPSS and its specific interface GHI to GPFS – a file system technology widely used in PRACE – has been selected as technology to be investigated, since it appears to be the most scalable system for handling huge amount of data. This is reflected by the fact that the biggest data centers in the world rely on this technology. Thus this seems to be also a technology providing the most advantages when considering efficient long term storage of data and therefore could be of interest for other PRACE partners beside RZG, where HPSS is implemented as archival system providing also an HSM functionality to the GPFS of the HPC-systems.

HPSS

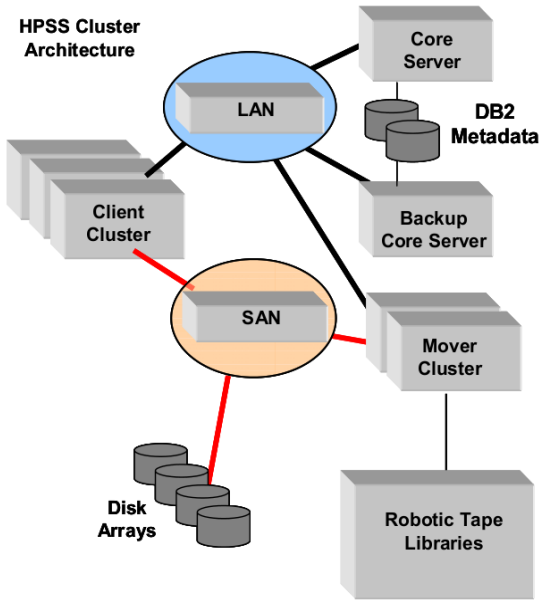
The High Performance Storage System (HPSS) is a software product developed through a partnership between US government labs (LLNL, LANL, LBNL, ORNL, SNL) and the IBM Corporation. It can store and retrieve large amounts of data on disk and tape libraries at high rates. The key to its high performance is its cluster design, combining multiple computer nodes into an integrated storage system. By increasing the number of nodes in the HPSS cluster, the system can be made to scale to any desired degree of performance. Current HPSS sites have dozens of petabytes in hundreds of millions of files and attain sustained data rates of several gigabytes per second.

The HPSS system consists of a Core Server, which manages the metadata using an IBM DB2 database, and any desired number of Data Movers, which have access to one or more disk cache systems and to one or more robotic tape libraries. The Core Server and the Data Movers are integrated in a high-speed network. Incoming data streams connect to an available Data Mover (transparently chosen by the Core Server), which ingests the data to its local disk cache and moves it (if desired, as assigned by its “class of service”) asynchronously to tape. When a client requests reading data, the Core Server determines where the data is located and instructs a Data Mover to read the data from tape or disk cache and to send it through the network to the requesting client.

Storage in the HPSS system is defined as one or more dynamic hierarchies, consisting of various combinations of low, medium and high speed devices. A so-called "class of service" is assigned to each stored file, allowing a flexible information lifecycle.

The HPSS core and movers run on AIX or Linux. The clients can be any architecture and access HPSS through one of several interfaces:

- Standard FTP
- Parallel FTP, a special HPSS implementation of FTP which allows several parallel data streams into or out of HPSS
- GridFTP
- Virtual File System (VFS). On Linux machines, the HPSS namespace can be mounted as a file system. Once mounted, the Linux machine can re-export the file system through NFS or Samba for access by other machines.
- Client API, a POSIX compliant interface to HPSS data for the application programmer.
- GHI, the GPFS-HPSS Interface (see below)



The cluster design of HPSS not only brings high performance, it also provides high reliability: if a Data Mover goes down for some reason, the system goes on working with the remaining Data Movers. If the Core Server goes down, any of the Data Movers can be promoted to Core Server in a short time by restoring a full backup of the Core Server. Optionally, the HPSS High Availability feature allows having a second Core Server in stand-by, becoming the active Core Server if the first one goes down.

Figure 1 aside shows the schematic modular configuration of an HPSS system.

Figure 1: Schematic setup for HPSS

A more detailed configuration showing also the data flows is presented in Figure 2.

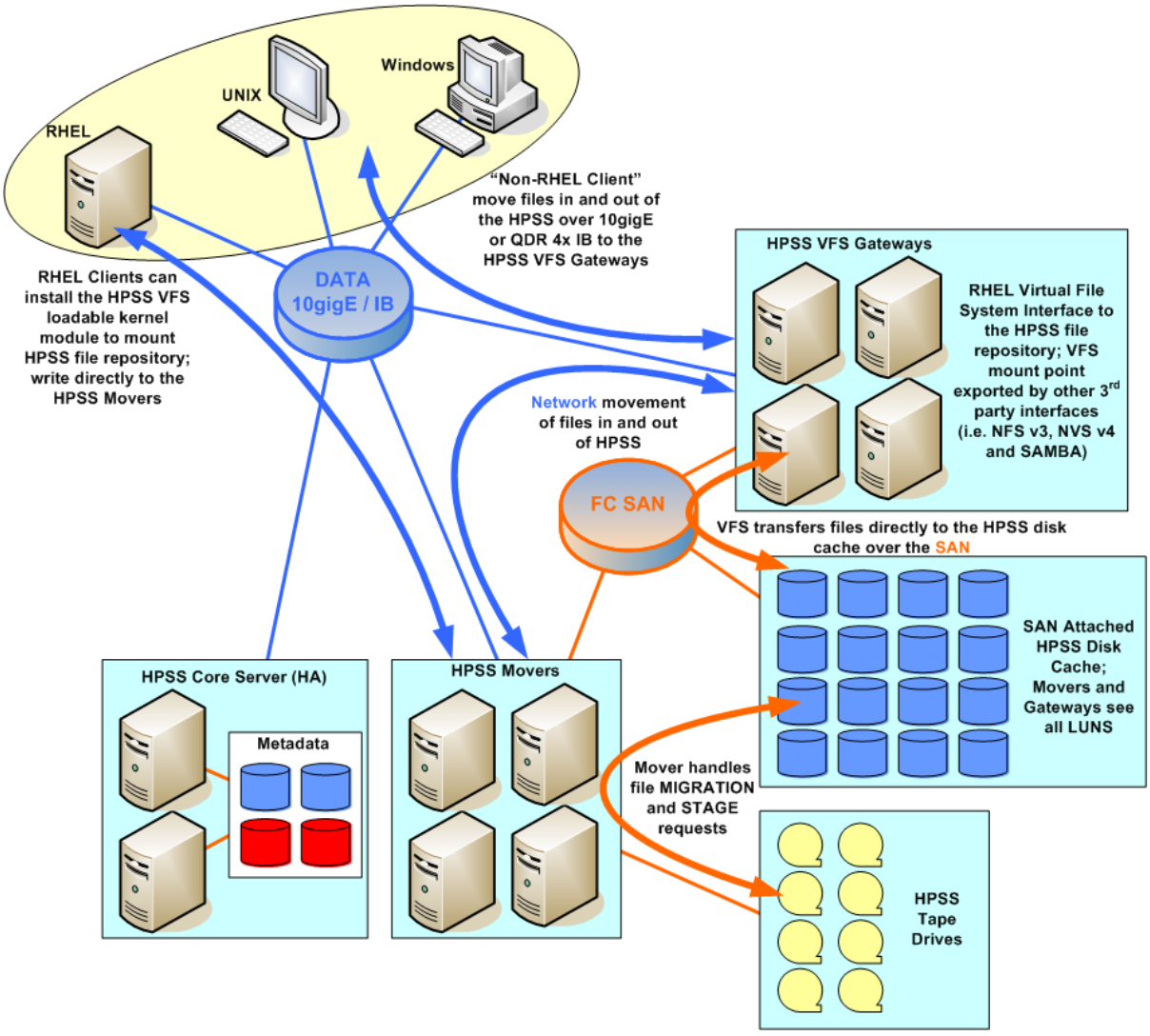
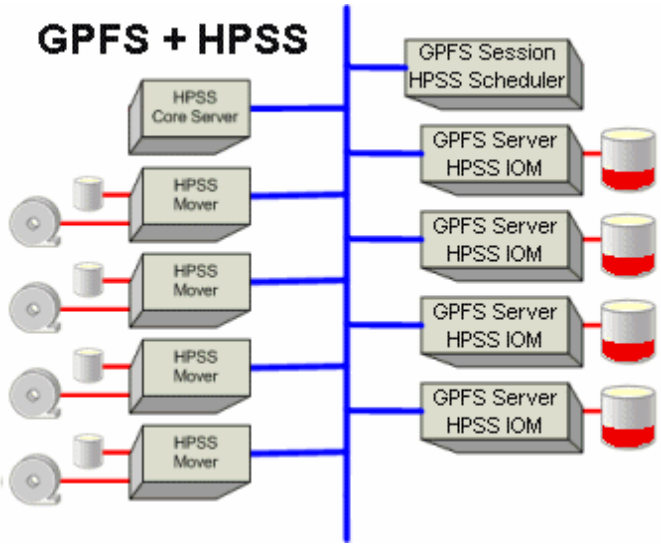


Figure 2: Detail configuration and data flow with HPSS

GHI

A special HPSS access mechanism is the GPFS-HPSS Interface (GHI). It connects GPFS and HPSS together under the GPFS Information Lifecycle Management (ILM) policy framework. GHI provides a Hierarchical Storage Management (HSM), offering the user a file system with virtually unlimited storage space. By defining appropriate GPFS ILM rules, the administrator can instruct GPFS to move files to HPSS when the file system gets full over a certain level. Users see all their files nevertheless, and if they want to read a file which has been transferred to HPSS, GHI transparently retrieves the file from HPSS and delivers it to the user.



GHI takes advantage of the multi-node capabilities of both GPFS and HPSS to move data in and out of HPSS using parallel connections, achieving very high data transfer rates as shown in Figure 3.

An added benefit of GHI is that it can store GPFS snapshots in HPSS, thereby providing disaster recovery protection for the GPFS file system.

Figure 3: GHI: GPFS interface to HPSS

Implementation at RZG

For testing the different functionalities provided by HPSS, a specific configuration has been setup at RZG. This consisted of

- 1 HPSS core server
- 4 HPSS movers with a total of 250TB cache
- GPFS with four servers and 35 TB diskspace
- 2 dedicated tape drives in the tape robot
- 20 dedicated tapes
- 10 Gbit/s network connectivity
- Linux-System for accessing HPSS via the different existing methods

For GPFS the several automatic migration configurations, depending on age, size, name, etc. have been successfully tested. Also automatic retrieving of migrated data was working as expected. The data transfer rate was dependent on whether the data was still in the HPSS cache or already migrated to tape. The mount-time for the tapes was below 1 minute. The transfer to the HPSS cache and then back to the GPFS was limited by the tape-stream and the network velocity. The test cases did not cover scaling by parallel reads from tape.

For non HPC systems the VFS interface implementing a POSIX view on the HPSS showed behaviour as expected for externally NFS mounted file systems, but this seems not a reasonable setup in connection with an HPC system. Similar results and limitations could be expected by accessing HPSS directly with FTP.

For more details of the evaluation, describing all tests performed and reporting their results see the appendix 5.4.

4 Remote Visualization

Task 10.3 “Remote Visualization” so far has focused on the following goals:

- To collect information about remote visualization solutions and services that have been evaluated and/or deployed by participating partners, including experiences and real world use cases, applications, issues, deployment experiences and user evaluation. (Sects. 4.2 and 4.3). This serves as the basis for working out general technology recommendations and best-practice guidelines for the PRACE project.
- To experiment with and assess new technologies and eventually develop new software components. To this end a first pilot project has been implemented at CINECA (Section 4.4).

Task activities have so far concentrated on “classical” remote visualization technologies (as opposed to localized “high-end” visualization devices such as power walls, caves etc.) which in essence are based on server-side rendering at the HPC centre and (thin) clients for handling data transfers and display at the end-user’s low-cost device (see Sect. 4.1 for details). This is considered as the natural starting point for establishing a distributed, trans-national HPC visualization service in the context of PRACE.

4.1 Introduction

The following Figure 4 sketches the components of a generic client-server system: A “Remote visualization” system specifically could be thought of as a client-server application where responsiveness and reduction of perceived latency are the most important requirements. In a general sense, remote visualization systems are used when the users have the need to interactively access data that is not directly available on their interaction device (workstation, laptop, tablet, and phone).

The alternative to any form of “remote visualization” is that the data to be visualized is transferred to the interaction device, first. This, however, is often unfeasible, either due to the lack of resources of the client (compute and storage capabilities) or because of prohibitively long transfer times given by the ratio between data size and effective network bandwidth.

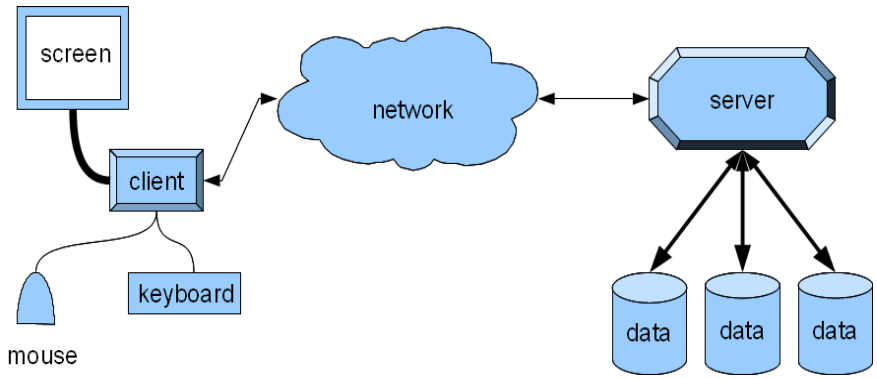


Figure 4: Generic configuration of a remote visualization system setup

Different strategies are being adopted for addressing this generic problem: the main issues are deciding on which kind of data is transferred via the network, on how to distribute processing between server and client components, and on how the total latency is distributed during the interaction process.

For example, in traditional web applications (Figure 5) the client is a browser and the server is web-server, the transferred data are html content pages, interaction is limited to html rendering, latency is concentrated under explicit user actions such as clicking and page load.

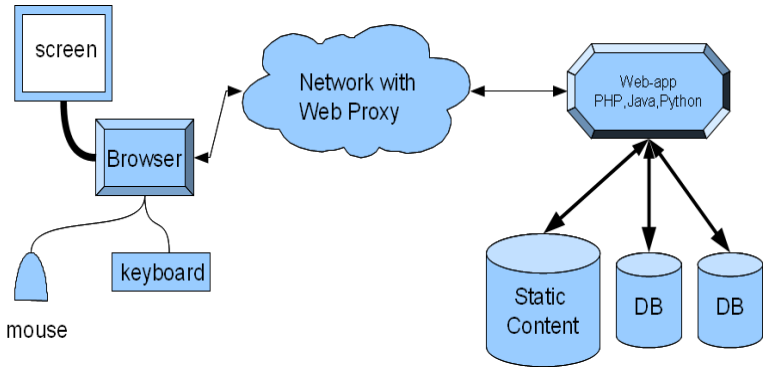


Figure 5: Components of a generic browser-based application setup

Several techniques have been developed to reduce latency and optimize bandwidth usage: caching (proxy and browser) techniques reduce latency for static contents, image compression optimizes bandwidth requirements, and incremental HTML rendering reduces perceived latency. More recently, web applications have adopted "AJAX" technology to provide a higher level of interactivity which relies on increased computing capability on the client side and implements complex interaction logic in JavaScript on the client.

In video streaming (Figure 6) and tele-conferencing (Figure 7) sequences of images are transferred over the network; applications, image and video compression are heavily used to better use available bandwidth; in video-streaming applications, aggressive pre-processing of video sequences that use several image frame buffer can be used. Interaction requirements, however, prevent these techniques to be applied in tele-conferencing applications.

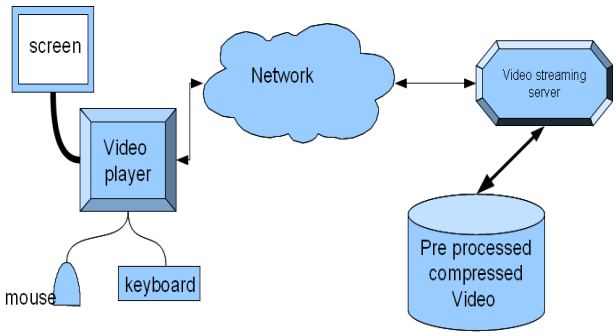


Figure 6: Video-streaming configuration

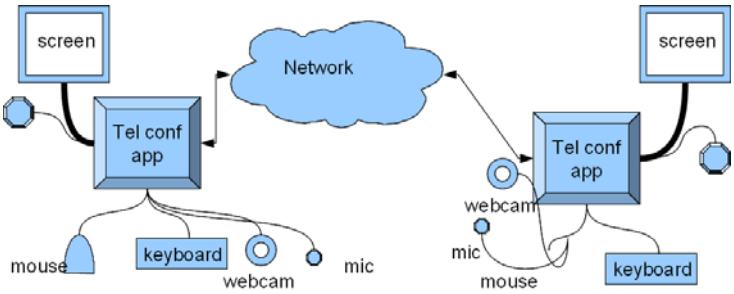


Figure 7: Video-streaming configuration

"Remote visualization" systems are commonly described as terminal-emulation frameworks such as VNC that allow to emulate the behaviour of a local session on a desktop workstation (the server) on a remote device (the client).If the connection is via WAN with limited bandwidth and/or the application uses hardware acceleration for 3D rendering, the simplistic approach of sending drawing commands over the network is obviously not feasible. Instead, remote rendering within virtual terminal frameworks has to adopt advanced image compression and transportation approaches. Usually, a compressed image stream is much more network-friendly than a stream of verbose graphics commands (like X11 GFX). As a side-effect of this concept of server-side rendering, also the application software (see below) can be installed and maintained centrally (i.e. at the HPC centre), thus removing the burden of software distribution and maintenance by the end-users.

A number of different software solutions for "remote visualization" have been developed for various platforms (operating systems)

Windows platform

- Windows Remote Desktop Connection
- Teamviewer
- Instant Housecall

Cross platform

- TightVNC
- Chrome Remote Desktop Extension

4.2 State-of-the-Art

This section describes available software solutions for supporting remote visualization, surveys their deployment at different partner sites and reports about specific activities and the status of a first pilot project.

4.2.1 *Scientific visualization applications supporting Client-Server paradigm*

While for general IT applications the remote desktop paradigm has already been widely used, large data sets arising in 3D visualization and analysis of massive scientific datasets pose additional challenges concerning bandwidth and latency optimization. Furthermore, in the scientific community the Windows operating system (OS) is not as widespread. The remote host system can be an HPC system or a visualization cluster; both are typically operated under some variant of the Linux or UNIX OS.

For these reasons, many of the scientific visualization applications are not just based on standalone software packages that leverage 3D acceleration but also require flexible client-server modality to allow users to install the client side of the application in their local desktop/workstations and to use highend host resources for interactively running the server-side of the visualization application. Some of these applications such as ParaView or VisIT are able to use cluster visualization resources for distributing the processing and visualization tasks among the available resources of the server cluster using MPI.

The visualization tools can be categorized as follows:

- General purpose visualization tools
  - VisIT [34]
  - ParaView [35]
  - IDL [36]
- Pre/Post processing tools for industrial simulation code
  - StarCCM [37]

4.2.2 *Application neutral, session oriented, VNC-like solutions*

This kind of solutions allows the remote usage of 3D, interactive applications in an almost transparent way, typically within a VNC-like, interactive session. The processing and graphics is carried on at the "host" side, the (thin) client sends input events from the local input devices (keyboard/mouse) and receives the image stream. This happens at a low level of the windowing interface, in a manner which is almost completely transparent to the different applications. The thin clients are often variations of the VNC-type of remote session handlers (RealVNC, TigerVNC, TurboVNC). The following Table 2 lists the most important remote visualization technologies and provides an overview of their implementation at PRACE partners.

Technology	Vendor	License	Links	Used in production
VirtualGL / TurboVNC	virtualgl project	LGPL v2.1	Homepage [20]	LRZ [21] RZG [22] SARA [23] CINECA (new)
RVN / DCV	IBM (discontinued)	Proprietary	Whitepaper [21]	CINECA (old)
ThinAnywhere	Mercury Intl.	Proprietary	Homepage [25]	unknown
NICE DCV (formerly IBM)	NICE	Proprietary	Product Description [26]	unknown
PCoIP	Teradici Corporation	Proprietary	Overview [27]	SNIC/LU
NVRemote (Monterey API)	Nvidia	Proprietary		Not yet available

Table 2: Remote Visualization technologies in use at the different PRACE sites

4.3 Existing Partner Visualization Services

Information has been collected in the task regarding the different visualization services which have been deployed by different partners.

Table 3 is a summary table of the resources partners will provide for supporting the activities of this task such as testing and evaluation of the remote visualization technologies and hosting pilot projects. Access to these infrastructures will be granted to the partners involved in this task as a first step. In a second step selected scientific users will be given access to these services. This will serve as the basis for working out recommendations and best-practice guidelines for the PRACE project.

Site	Public Viz	Nodes	GPU	Memory	Reserved Viz	Nodes	GPU	Memory	High end local viz systems
LRZ	no	n/a	n/a	n/a	yes, reserved for LRZ Linux cluster users	2 servers	2 Nvidia Quadro FX5800 cards per server	256GB per server	5-sided cave, 4K 3D powerwall
RZG	yes	5 render nodes: 2x Intel W5580 4c, 1 'fat' render node: 4x Intel X7542 6c, 1 login node	NVidia FX5800 (2 per node)	5x 144GB, 1x 256 GB ('fat' node)	n/a	n/a	n/a	n/a	2 workstations with active stereo (120 Hz monitors and NVidia shutter glasses)
SARA	yes	16 render nodes + 1 login node: Xeon E5620, 800 GByte scratch	NVidia Gefore GTX460	12 GB	n/a	n/a	n/a	n/a	n/a
SNIC/LU	Campus wide PCoIP Teradici	2 dedicated viz nodes with PcoIP hardware in current test	Nvidia Quadro 5000	32 GB per viz node	yes (P2P)	2 viz (208 compute)	n/a	n/a	n/a
CINECA	Ongoing deployment, VIZ PBS queue	2 out of 274 double esa-core Intel Xeoncompute nodes on a Viz dedicated queue	2 Nvidia Tesla M2070	48 GB	in production for industrial users	6 double quad core Intel Xeon	2 Nvidia Quadroplex 2200 S4	128 GB	curved screen VR room

Table 3: Existing partners' visualization services

It is apparent that currently VirtualGL+TurboVNC is the most widely used technological platform for remote visualization. This solution was presented in detail by SARA during the CINECA summer school on scientific visualization (see below and the presentation at the Summer School [19]).

4.4 CINECA pilot project: GUI manager for a remote visualization TurboVNC session using PBX job scheduler.

Along with the PRACE activity, CINECA has decided to migrate from DCV technology to VirtualGL+TurboVNC for the deployment of a more scalable visualization service.

Due to the presence of graphics hardware on each node of the cluster, the most scalable approach would have been to provide visualization resources from the general-purpose cluster pool of compute nodes (more than 256 potentially available). Unfortunately, this decision

requires substantially higher efforts by the user for setting up the remote connection (ssh tunnel). According to preliminary experiments with the CINECA user base, especially from the technical computing area, it has become quite clear that an improvement in the setup procedure is a crucial point for a broader adoption of the service. To this end, a new software component has been developed which facilitates the TurboVNC installation, setup and connection management for the user. While this project has originally been tailored towards the CINECA deployment environment, it could well be adapted to different environments in the PRACE context. A detailed description of the application is given in the appendix 5.5.

A number of CINECA users are already testing this setup. It is planned, in a first step to extend the CINECA user base by end of year and also provide PRACE partners an account to let them remotely test this service. Depending on the feedback and requirements of other partners that have similar technology in place (SARA, RZG) the software can be adapted to fit other site's requirements such as a different scheduler or different access methods. It will then be tested on another infrastructure.

4.5 *CINECA Summer school of Scientific Visualization*

In June 2012 CINECA has organized a summer school of scientific visualization [28] as an intense, 5-day, graduate level course.

Topics of this course have been: (see [29] for the detailed program and course material):

- Introduction to computer graphics
- Introduction to scientific and remote visualization
- High-performance visualization tools and libraries
- Advanced techniques for in-situ visualization
- Acceleration engines for complex scenarios

Specifically, the PRACE partners RZG, SARA and CINECA have presented their visualization infrastructure and services along with a number of general visualization topics:

- Introduction to VisIT and RZG/MPG remote visualization services (RZG)
- Remote visualization with VirtualGL (SARA)
- In situ visualization with ICARUS and ParaView (CINECA)

In addition, the experimental CINECA visualization service and connection manager has been used to support school exercises.

5 Annex

In the Annex more detailed information for some of the subtasks is collected.

5.1 Storage Accounting Questionnaire

'Storage accounting and disk usage information for users' internal survey aims to gather information from all partners and AISBL to discuss the motivations and needs of the disk usage accounting.

This survey has 27 questions and is divided into three sub-sections. Below is the list of the sub-headings:

- Current status of sites
- Policies and requirements
- Implementation related questions

Methodology:

Open-ended questions were created for this survey. The reasoning for the open-ended questions is to collect the current opinions of all the partners and AISBL on the storage accounting subject.

CURRENT STATUS OF SITES

1) How many users and projects are there on your site?
2) What types of storage systems are there (SSD, SAS, SATA, tape…)? What size are they?
3) How much space is used on your storage systems?
4) What is the average space usage of a user and project on your storage system?
5) Which storage classes have been implemented on your storage systems? (archival, permanent, temporary, backed up/not backed up, etc…)
6) What is your storage planning strategy?
7) What kind of storage utilization policies does your site have?
8) What kind of storage accounting policies does your site have?
9) What kind of storage usage information is being provided to users and funding agencies by your site?
10) What tools are being used for getting disk usage information (du utility, standard linux quota utility, any other third party tools, etc…)?
11) Are there any reporting tools for disk usage information on your site?

POLICIES AND REQUIREMENTS

12) What could be the motivation of users and/or sites to record the storage accounting data?
13) What kind of storage utilization policies should be implemented?
14) What kind of storage accounting policies should be implemented?
15) What kind of reporting tool would be appropriate for storage accounting records?
16) Would the storage accounting information help the storage planning process?

IMPLEMENTATION RELATED QUESTIONS

Some of the questions of this part of survey are related to European Middleware Initiative's Storage Accounting Record (EMI STaR) proposal [8].

17) What should be the granularity in time of the storage accounting record? (accounting record frequency, time frame of validity of a record, time-stamp type of the records, etc…)
18) Which technical barriers and/or problems would be encountered while recording fine-grained storage accounting data in terms of file system administration? (e.g. It's a very CPU intensive process for meta data server of Lustre FS to record all the usage activity hourly.)
19) What would be the accounting unit for disk usage records? (GB-hours, or just GB)
20) Should the file access counts and/or bandwidth usage for accessing the file be reflected on the accounting records? What kind of problems could reveal this kind of accounting records, if any?
21) Which space should be accounted? The reserved space or the used space?
22) Should different type of storage systems be distinguished in terms of accounting records? Should different coefficient values be used while billing for the different types of storage? Should billing and accounting records be separated?
23) Should the storage accounting be based on user/projects or on files?
24) Would there be any need for central storage accounting system? Should Apache – DART system also be considered for storage accounting?
25) Which mechanisms would be needed for exchanging the accounting records between sites securely and reliably?
26) Could service level agreements be part of a storage accounting record? What would be the way for including SLAs into accounting records?
27) What kind of functionalities should provide a reporting tool for storage accounting?

5.2 Comparison of HPC-Europe and CINES tools for proposal management

#	Functionality	Rate	HPC-Europa	CINES	Comments
1	Electronic submission of project proposal.	Essential	Y	Y	
2	Developers ability to programmatically redesign the forms contents and their integration with the internal database.	Essential	Y	Y	
3	Web-based ability (form design tool) to design and change the project submission and evaluation forms.	Desideratum	Y	P	CINES : Planned in portable kernel roadmap.
4	Provide users with complete online control of their data	Essential	Y	Y	

	(application form, user data etc.) and enable them to effectively view and browse their data (i.e. applicants can see all their applications, response letters and applications status form the portal).				
5	Assign different roles (coordinator of the process, evaluator etc.) and give access to different functionalities (i.e. evaluation assignment, evaluation process), views and data (statistical, project submission form and evaluation form) according to the different privilege level (i.e. evaluators can gain limited access to relevant proposals and TE). This would cause different log-in views for Applicants, Technical & Scientific evaluators and DAAC staff.	Essential	N	Y	CINES : Roles are implemented. However, Admin UI delegation is not yet available. HPC-Europa : no web-based authorization management is currently provided. Heavy changes to the kernel are required.
6	Store applicants' data, project data, TE review data, suggested extra TE info, SE data, ranking info etc. into the DECI Database.	Essential	N	Y	HPC-Europa : Integration with the DECI database is needed.
7	Create and/or change user's, evaluator's, site's, countries, info.	Essential	P	P	CINES : Planned 2H2012 and/or portable kernel roadmap HPC-Europa : evaluators cannot change their info autonomously while users can.
8	Support the process of submitting a short report from the PI, after the completion of the project; the template of this report being downloadable from the tool.	Essential	Y	Y	
9	Create statistics reports of the DECI process (i.e. number of technical evaluations per site, number of scientific evaluations per evaluator). Moreover the publications related to work done with DECI resources should be	Desideratum	P	Y	CINES : Should be ok, to be precised. HPC-Europa : general statistics on the entire review process are

	tracked via the proposed tool.				available though.
10	Copy or link the relevant data from the web-based tool, when needed, into the DPMDB (i.e. project name, home site, technical requirements such as CPU type, number of jobs, memory, simulation codes etc.).	Essential	N	P	CINES : Linking should be possible, with quite reasonnable work HPC-Europa : Integration with the DPMDB is needed.
11	Copy summary of projects' resource usage from DPMDB to the web-based tool, so that PIs can view accounting information related to their projects without learning a new tool (DART).	Desideratum	N	P	CINES : Should be possible
12	Create and export documents and information that should feed other systems or processes (i.e. automatic generation and export of PDF's for mailing at any point in time). Enable generic export (all documents related to a call to be exportable in corresponding folders/files - e.g. one folder "Astrophysics" containing as many as folders as proposals, each containing all the documents related to this proposal = application + tech review + scientific review)	Desideratum	P	Y	CINES : Such features already exists in PPR tool, but some specific development may be necessary to fit the requirements. HPC-Europa : Most of the information can be easily exported via Excel file format, nor PDF.
13	Keep extensive logs regarding all changes made by the users in the tool.	Desideratum	Y	Y	
14	Provide different communication tools (via email, via user workspace etc.) between the users who have to communicate according to the existing workflow (i.e. technical evaluator and principal investigator).	Desideratum	Y	Y	
15	Design and run workflows between the Coordinators of the Evaluation Process, the evaluation sites and the evaluators. The web-based DECI tool could support rule creations that would be	Desideratum	N	P	CINES : Included in kernel development roadmap. HPC-Europa : Easy to

	associated with conditions and actions (i.e. time reminders or enforcement – establish deadlines for submission of evaluation, email reminders to reviewers, alerts to the evaluators of completed, pending or overdue reviews).				develop.
16	Provide administrator with complete autonomous control of the tool parameters - e.g. reopening applications (needed in the administrative process), changing the deadline of a review, changing the discipline category of a project (when the automatic categorization failed)	Essential	P	-	HPC-Europa : Basic tools (e.g. reopening, deadline change, etc.) are already available. Advanced ones should be better clarified.
17	Communicate to the centers the info of awarded projects (LDAP) "Project ID, User Accounts, etc."	Essential	P	-	HPC-Europa : LDAP compliant information can be already exported but specific developments could be necessary according to LDAP schema.
18	Create a report of reviewers, with past historical information (reviews attributed and reviews in previous calls), including passwords	Essential	P	-	HPC-Europa : Easy to implement.
19	Create a report of all persons involved in past and present calls (PIs, collaborators) with history (call, proposal ID, ...)	Essential	Y	-	
20	Guarantee a highly secure log-in system (highly secure password)	Essential	Y	-	

Legend

Y = well supported

P = partially supported

N = not supported

- = not ranked yet

5.3 iRODS-Workshop Preliminary Agenda

Wednesday, September 26

Session 1: Introduction to data management

13:30 **Workshop opening**
Agnès Ansari, CNRS/IDRIS
13:40 **iRODS in Sweden**
Tom Langborg, SNIC/LIU
14:10 **Introduction to iRODS**
Leesa Brieger, DICE

15:10 Coffee break

15:30 **Demo of basic capabilities and hands-on training**
Leesa Brieger (Jean-Yves Nief, Agnès Ansari)

17:30 End of day

Thursday, September 27

Session 2: iRODS tutorial

09:00 **Introduction to rules and micro-services**
Leesa Brieger, DICE
09:30 **Simple rules and data base queries**
Leesa Brieger, DICE
10:00 **Complex rules and scheduling**
Leesa Brieger, DICE

10:30 Coffee break

Session 3: iRODS applications

11:00 **iRODS at CINES**
Gerard Gil, CINES
11:20 **iRODS at CC-IN2P3**
Jean-Yves Nief, CNRS/IN2P3
11:50 **iRODS status in Sweden**
NN

12:30 Lunch

Session 4: Users needs and requirements

13:30 **iRODS experience in EUDAT**
Giuseppe Fiameni, CINECA
14:00 **iRODS experience in DEISA**
Agnès Ansari, CNRS/IDRIS

14:25 **Tiers 0 – Users needs and requirements**
Stefanie Janetzko, FZJ
14:50 **Tiers 1 – Users needs and requirements**
Chandan Basu, SNIC/LIU

15:15 **Discussion**
Jean-Yves Nief, CNRS/IN2P3

15:45 Coffee break

15:30 **Demo and hands-on training**
Leesa Brieger (Jean-Yves Nief, Agnès Ansari)

17:30 End of day

19:00 Dinner

Friday, September 28

Session 5: Advanced topics

09:00 **Authentication and Authorization in a federated environment**
Jules Wolfrat, SARA
09:30 **iRODS security**
Reagan Moore, DICE
10:00 **iRODS performance**
Reagan Moore, DICE

Session 6: Strategy and future

10:30 **Users requirements and needs summary**
Agnès Ansari, CNRS/IDRIS

10:40 Coffee break

11:10 **iRODS strategy and future**
Reagan Moore, DICE
11:40 **DAITF and the DataNet Federation Consortium**
Reagan Moore, DICE

12:10 **Discussion and wrap-up**
Tom Langborg, SNIC/LIU

13:00 End of the workshop

5.4 Functionality, Performance and Failover/Recovery of HPSS and GHI

For HPSS and GHI several tests have been performed to verify basic functionality, to evaluate performance and to check failover and recovery abilities. In the next six sub-sections these 27 tests are shortly described and their general results documented.

5.4.1 *HPSS Functionality Tests*

1) Reading and writing files into the archive system through FTP
   - Linux machine connected with the standard ftp command to HPSS and tested writing and reading some files to/from HPSS successfully.
2) When writing files through FTP, setting different Classes of Service for different files
   - Used 'quote site setcos' command successfully to test specifying a Class of Service.
3) Reading and writing files through parallel FTP
   - After installation of a PFTP client on several machines sharing a GPFS file-system the PFTP multi-node capability to transfer files to/from HPSS was used. It could be verified that all participating nodes have been transferring data (with "iostat").
4) Concurrent write access to multiple tapes and concurrent read access from multiple tapes
   - Two Classes of Service defined sending incoming data directly to tape. Then from two Linux machines data was written via ftp using both classes of services and it was verified that at least two tapes got mounted and written to at the same time. After the write process had finished and the tapes got dismounted, the same data was read back and it was verified that at least two tapes got mounted and read at the same time.
5) Test API with HPSS sample code
   - The HPSS API sample programs have been compiled and the functionality was verified.
6) Ability to run user authentication for all HPSS operations with external Kerberos
   - For all operations above, a userid/password had been defined in a non-HPSS Kerberos-domain and this userid/password has been successfully used for operations with the HPSS.
7) Kerberos password-free authentication (with GSSAPI)
   - Similar tests as with userid/password authentication have been successfully performed for test 3), which was the only tool supporting GSSAPI as authentication service, using Kerberos credentials, which proved functional without being required to enter any password.

5.4.2 *HPSS Performance Tests*

8) Aggregated read/write transfer to/from disk cache
   - The PFTP nodes configured for test 3) have been used to do read/writes to HPSS. The transfer-rates have been measured. It turned out that with 4 PFTP nodes a combined throughput above 2 GB/s for a period of at least 30 minutes could be achieved.
9) Many (100) parallel recalls at the same time via ftp with files already in the disk cache
   - Created scripts for controlling a parallelized workflow and system monitoring to check swapping behaviour comparing 1, 10 and 100 recalls.
   - System stays operational, no lock situation appears, no crash, no huge performance degradation, login still possible with reasonable response times.

- On up to 4 nodes several normal (not parallel) FTP file reads at the same time have been started with the total number of recalls of 1, 10, and 100. On one of the four mover-nodes the swap space was monitored by writing the output of 'vmstat 1' to a file. During that operation other simple operations, like logging in as admin and querying some parameters have been performed and it was verified that the system remained usable.

10) Many (100) recalls at the same time via ftp while all files had to be loaded from tape
    - Created scripts for controlling a parallelized workflow and system monitoring to check swapping behaviour comparing 1, 10 and 100 recalls.
    - System stays operational, no lock situation appears, no crash, no huge performance degradation, login still possible with reasonable response times.
    - The test was in principle identical to the previous one, while just in addition a tape fetch had to be initiated.

5.4.3 *HPSS Failover and Recovery Tests*

11) Handling of an outage of a tape drive or when a tape get stuck in the drive (simulated by switching off the tape device)
    - HPSS does not break on switch off, but on read the data is not automatically loaded from the second copy. The outage is transparent to HPSS users in that the read-call stays open without any error-message.
    - After switching on the tape-device again, for freeing the tape a manual intervention was required on the system controlling the robot-system.
12) Simulation of an outage of one HPSS-metadata disk and one disk cache disk by drawing out a disk out of the storage system
    - HPSS does not break while removing a disk during a ftp-recall of a file resident in the HPSS-cache. The outage is transparent to HPSS users and has no impact on the functionality; only the performance is degraded.
13) Crash DB2-database of HPSS and recover from backup-version
    - Created some files managed by HPSS, backup the HPSS DB2-database with Tivoli Backup. Simulated a database-crash by deleting some of the DB2-volumes while the system is running.
    - HPSS went down. After restoring the system including the database from the backup with Tivoli Bare Metal Recovery (TBMR), HPSS could be started again.
    - Consistency and correctness was verified by reading back the files written in the first step.
14) Outage and migration of the core-server functionality to a mover-machine
    - Switching off the core-server crashed HPSS.
    - After restoring the core-server on a mover-machine with TMBR a restart of HPSS was successfully performed. All functionality was available again in less than 4 hours.
15) Outage and bare metal recovery of a mover
    - Switching off a mover caused access to data located on that machine to be interrupted, but HPSS in total continued to be available.
    - After restoring the mover with TMBR full functionality was recovered.
16) HPSS functionality and performance can be monitored through Nagios
    - Several monitoring functions have been successfully implemented.
17) Emulating a failure of the database controlling the tape-robot:
    - In one case the tape-robot-control-machine has been stopped and in the other one the network connectivity has been disrupted. After a file-recall with ftp no error message has been shown up. The process stayed hanging.

5.4.4 *GHI Functionality Tests*

18) Transparently access migrated files in the GPFS
    - After writing some files to a new GHI-controlled GPFS file-system appropriate HPSS admin commands have been applied to get the files copied to tape. Then those files have been read again from the file-system successfully after retrieving them transparently from tape.
19) Running migrations once, migrate files from GPFS to HPSS
    - After storing a file the GHI watermarks have been lowered to force migration. The successful migration was checked by comparing the residency flags of that file.
20) Triggering purges through high/low watermark policies
    - Writing sufficient enough data to pass the defined watermark caused migration and purging of the files according to the rules defined in the policies. The check was done by observing the residency flags of the files.
21) Check that pinned files in GPFS stay resident
    - Files can be forced to stay online, although a tape copy is generated with the command 'ghi_pin'. Trying to purge such files is then not possible. This was proven to be working as expected.
22) Verifying the staging functionality (retrieve from HPSS synchronous to user request):
    - With the command 'ghi_stage' purged could be brought online again, which was verified with command 'ghi_ls'. The content of the restored files was compared to the initial files and was verified to be identical.
23) Class of Service selection based on GPFS ILM policies.
    - Several different policies (size, age, etc.) have been tested successfully.
24) Remove a file and all backups that reference that file (garbage collection)
    - Files deleted from GPFS disappeared from HPSS after some time.

5.4.5 *GHI Performance Tests*

25) Compare recalling a file within GHI to reading a file with ftp from HPSS
    - Writing large files (1GB – 100GB) to a GPFS with 4 servers showed that the transfer-rates are comparable to those of PFTP with 4 nodes as done in test 3) within a range of 5%.
26) Compare file migration with GHI with writing a file via ftp into HPSS
    - Similar as previous test but for writing: the performance of GHI is also within a 5% variance to native HPSS.

5.4.6 *GHI Failover and Recovery Tests*

27) Backup and Restore of GPFS Metadata
    - All files from in the GPFS have been forced to be migrated into HPSS. Then the GPFS metadata has been backed up using "ghi_backup".
    - GPFS has been destroyed and then the saved GPFS metadata was restored using "ghi_restore".
    - After this restore of GPFS metadata all migrated files have been recalled using "ghi_stage".
    - After successful recall "ghi_ls" shows "B" as residency flag for recalled files and all files could be opened successfully.

5.5 Remote Visualization Pilot Project at CINECA

The Visualization Pilot Project at CINECA is the first pilot implementation as described in the DoW.

GUI manager for a remote visualization TurboVNC session using PBS job scheduler

5.5.1 *Requirements*

Users (academic and industrial) would like to perform **scientific visualization** on **large data sets** produced on CINECA HPC systems, offer an **high performance** environment even for visualization, possibly **without moving the data** from where it has been generated, easily.
HPC center would like to set up a service which is:

- **Scalable:** few resources could be initially allocated to the service but more could be added in later if needed.
- **Accountable:** the use of the services could be evaluated and eventually accounted to users
- **Reliable:** service should have the same reliability as the other computing center resources

5.5.2 *Allocated resources and deployment constraints*

HARDWARE:

- 2 PLX compute nodes with GPU
  - Processors: 2 six-cores Intel Westmere 2.40 GHz per node
  - GPU: 2 NVIDIA Tesla M2070Q per node
  - RAM: 48 GB/node

- RVN05 (inbound connection and login allowed):
  - Processors: Quad-core Nehalem IBM E5540 Intel(R) Xeon(R) CPU at 2.53GHz
  - GPU: 2 NVIDIA Quadro FX 1700
  - RAM: 128 GB/node

SOFTWARE:

- Common home and scratch GPFS mounted by login, compute nodes and RVN node
- Module tcl environment system
- access only by PBS scheduler

NETWORK SETUP:

- INFINIBAND connections with all nodes of the cluster
- Inbound connection from outside not allowed on the two PLX internal nodes, allowed on the RVN node: This means that a ssh tunnel is required on the two PLX nodes for supporting any client-server external connection.
- Outbound allowed

5.5.3 *Remote visualization layer*

- **TurboVNC**: open source VNC client (remote control software) that support VirtualGL

- **VirtualGL**: open source package that gives any Unix or Linux remote display software the ability to run OpenGL applications with full 3D hardware acceleration.

Optimize user experience of remote 3D applications by rendering on remote GPU while streaming only the 2D result images.

5.5.4 *Deployment setup*

Layout of the components:

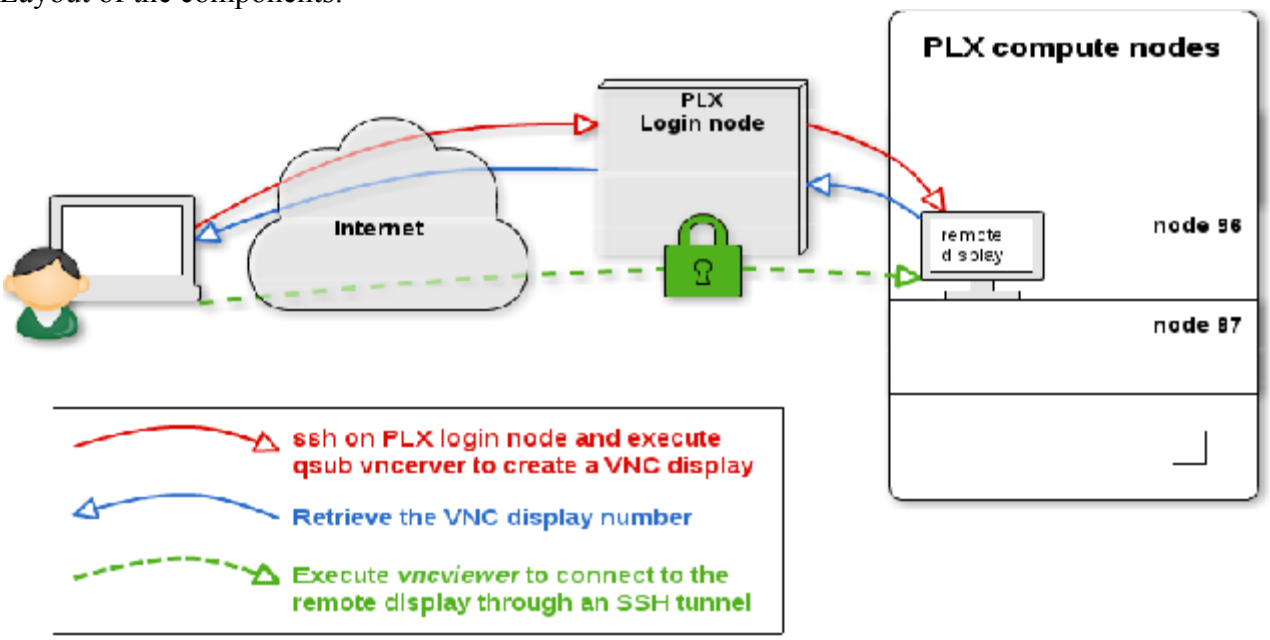


Figure 8: Schematic login process for starting remote visualization

5.5.5 *Remote Connection Manager*

We have developed a python cross platform application (tested on Windows, linux and mac OS) that simplifies and automates the steps needed for setting up theVNC connection to the visualization compute nodes (job submission for VNC server start, ssh tunneling, vnc client connection) and managing it (reconnection, list, close).
It is a client/server application that allows the user, through a user interface, to create remote displays, connect the ones he has created and kill the ones he doesn't want to use any more.

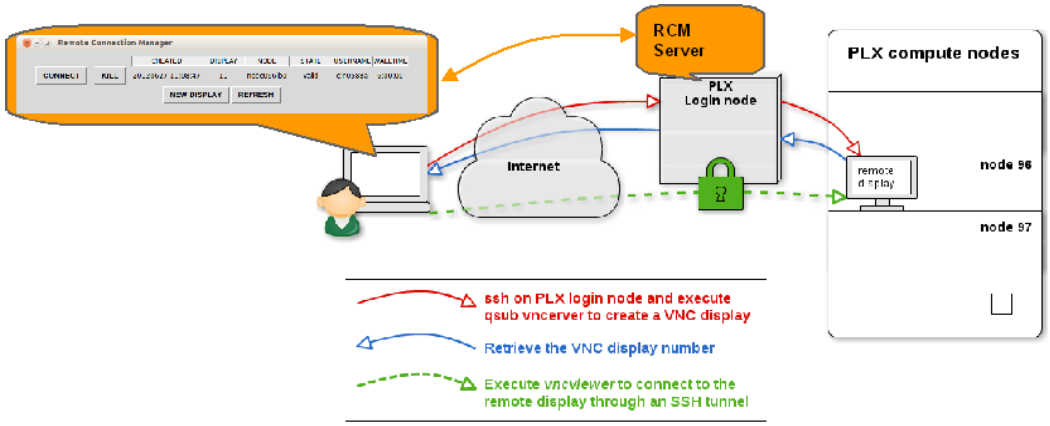


Figure 9: Login via Remote Connection Manager

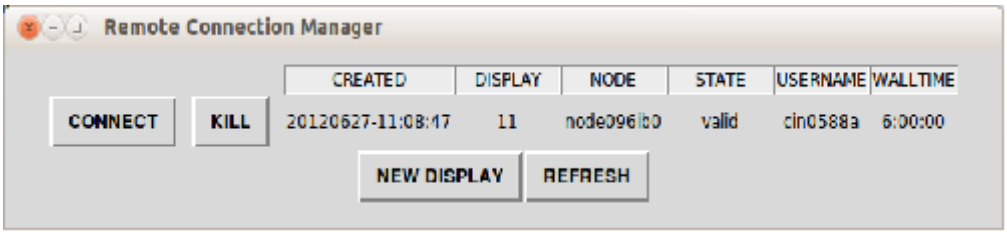
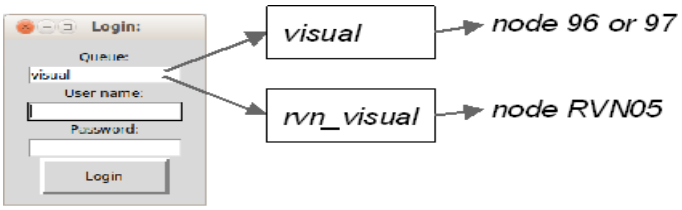


Figure 10: Main Panel for Remote Connection Manager

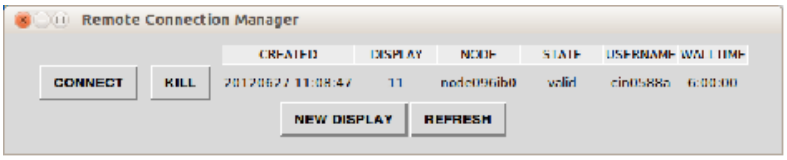
- On start-up and "**REFRESH**":
  - Server executes the command *qstat* to get the list of VNC instances available (running PBS job) and related information (display number, job-id, elapsed time)
- "**NEW DISPLAY**":
  - Server submits a PBS job to execute *vncserver* and stores the number of the VNC display created, then the Client starts connection (using one time password)
- "**CONNECT**":
  - Client runs *vncviewer* through an SSH tunnel specifying the VNC display number (on windows it requires user password)
- "**KILL**":
  - Server executes the command *qdel* to kill the VNC instance of the PBS job

There are three interaction steps:

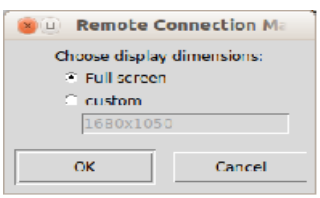
1. Insert queue and user credential



2. Create a new remote display (or connect to an existing one)



3. Choose remote display dimensions



5.5.6 *Deployment on PLX cluster*

- **2 PBS Queues:**
  - 1 queue on standard compute nodes, currently two, only difference is the requirement of X11 server running on each node
    - default queue for standard users (nodes with 48 GB each)
    - accounting not currently activated
    - users must currently explicitly ask for access to the experimental service and be authorized
  - 1 queue on "FAT RVN nodes" rvn_(one node of the 6 available):
    - queue reserved to specific users who need more RAM (node with 128 GB)
    - users need specific authorization
    - users can directly access from the external internet without ssh tunneling, PBS jobs not strictly required
- **Jobs WallTime set to 6 hours**.
  - the visual job session is automatically started but can currently run any task without limitation, when time limit is reached, all processes are killed
- **CPU "overbooking"**:
  - interactive vizualization applications produce a highly variable workload (resources are usually needed just when the user is active on the interface)
  - the PBS visual queue parameter "available.cpu" has been defined as a multiple (double) of real cores (24 on a 12 core node)

5.5.7 *Evaluation and further development*

- **Current evaluation**:
  - experimentation with current remote visualization users
  - collect feedback and usage statistics
  - testing with visualization applications
    - ParaView
    - Starccm
    - Blender
  - fine-tuning of the service
    - bug fixing
    - cross platform testing
      - Windows XP, Windows 7
      - OSX (Lion)
      - Linux boxes
        - ubuntu 32
        - ubuntu 64
        - RedHat EnterpriseLinux (RHEL)
    - small GUI enhancements:
      - progress bar to aware user of server side waiting (job creation, removal)
      - warn users of forthcoming session expiration
- **Forthcoming actions** (October until end of the year):
  - promote the service and open to other users

- possible increment the nodes allocated to the queue
- advanced reservation mechanism for specific usage patterns
- extending the tested applications
- defining a suitable accounting parameter (the compute nodes are currently accounted for elapsed time and this does not fit well with highly variable loads generated by interactive visualization applications)
- interaction with Tier 0 machine (Fermi Blue Gene Q)

The software of the Remote Connection Manager can be retrieved from an SVN-Server located at CINECA [30]. The respective documentation can also be found at CINECA [31].