



**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 Phase Project

Grant Agreement Number: RI-283493

**D10.2
Second Annual Report of WP10
Progress on Technology Scouting and Development**

Final

Version: 1.0
Author(s): Andreas Schott, GCS/MPG-RZG
Date: 23.08.2013

Project and Deliverable Information Sheet

PRACE Project	Project Ref. №: RI-283493	
	Project Title: PRACE Second Implementation Phase Project	
	Project Web Site: http://www.prace-project.eu	
	Deliverable ID: < D10.2 >	
	Deliverable Nature: <DOC_TYPE: Report / Other>	
	Deliverable Level: PU *	Contractual Date of Delivery: 31 / August / 2013
		Actual Date of Delivery: 30 / August / 2013
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: Second Annual Report of WP10	
	ID: D10.2	
	Version: <1.0>	Status: <i>Final</i>
	Available at: http://www.prace-project.eu	
	Software Tool: Microsoft Word 2007	
	File(s): D10.2.docx	
Authorship	Written by:	Andreas Schott, GCS/MPG-RZG
	Contributors:	Jules Wolfrat (SURFsara), Luigi Calori (CINECA), Gabriele Carteni (BSC), Agnes Ansari (CNRS/IDRIS), Tom Langborg (SNIC/LIU), Tilo Eißler (GCS/LRZ), Giuseppe Fiameni (CINECA), Ilya Saverchenko (GCS/LRZ), Miroslav Kupczyk (PSNC)
	Reviewed by:	Guillermo Aguirre (BSC), Dietmar Erwin (PMO)
	Approved by:	MB/TB

Document Status Sheet

Version	Date	Status	Comments
0.1	20/July/2013	Draft	Initial Version
0.2	23/July/2013	Draft	File System Technologies
0.3	25/July/2013	Draft	Service Certifications
0.4	26/July/2013	Draft	Accounting
0.5	26/July/2013	Draft	Event Integration
0.6	29/July/2013	Draft	Data Strategy
0.7	29/July/2013	Draft	iRODS

0.8	29/July/2013	Draft	File Transfer Technologies
0.9	30/July/2013	Draft	Formatting
0.10	31/July/2013	Draft	Reworking Contents
0.11	05/August/2013	Draft	Reworking Contents
0.12	06/August/2013	Draft	Monitoring, Annexes
0.13	07/August/2013	Draft	DECI-Portal
0.14	08/August/2013	Draft	Reworking Contents
0.15	09/August/2013	Draft	PRACE Information Portal
0.16	10/August/2013	Draft	Collaborations
0.17	12/August/2013	Draft	Remote Vizualisation
0.18	13/August/2013	Draft	Cleanup for internal review
0.19	16/August/2013	Draft	Reworking towards review comments
0.20	16/August/2013	Draft	Added iRODS Evaluations to Annex at the very end
0.21	17/August/2013	Draft	Moved back FS-performance from annex to main document
0.22	18/August/2013	Draft	Reworking towards review comments
0.23	19/August/2013	Draft	Bartosz for 2.3; work towards review comments
0.24	20/August/2013	Draft	Josip for 2.2; Ilya, Matteo and Jules for 2.7; Jules for 2.1; Anders for 4.3; Dietmar for 6.6; more formatting for 6.6
0.25	21/August/2013	Draft	Zoltan for 3.4; Luigi for 4; deleting unused acronyms; updating 2.7 with input from Giuseppe; extended appendix 6.4 for clarification of 3.2;
0.26	22/August/2013	Draft	Additions to PIP 2.5; reworked iRODS 3.3; reworked File Systems 3.4
0.27	23/August/2013	Draft	Finalizing Introduction and Summary; final review, corrections, and fine tuning formatting
1.0	23/August/2013	Final version	

Document Keywords

Keywords:	PRACE, HPC, Research Infrastructure
------------------	-------------------------------------

Disclaimer

This deliverable has been prepared by the responsible Work Package of the Project in accordance with the Consortium Agreement and the Grant Agreement n° RI-283493. It solely reflects the opinion of the parties to such agreements on a collective basis in the context of the Project and to the extent foreseen in such agreements. Please note that even though all participants to the Project are members of PRACE AISBL, this deliverable has not been approved by the Council of PRACE AISBL and therefore does not emanate from it nor should it be considered to reflect PRACE AISBL's individual opinion.

Copyright notices

© 2013 PRACE Consortium Partners. All rights reserved. This document is a project document of the PRACE project. All contents are reserved by default and may not be disclosed to third parties without the written consent of the PRACE partners, except as mandated by the European Commission contract RI-283493 for reviewing and dissemination purposes.

All trademarks and other rights on third party products mentioned in this document are acknowledged as own by the respective holders.

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.....	vi
References and Applicable Documents	vi
List of Acronyms and Abbreviations.....	vii
Executive Summary	1
1 Introduction	2
2 Enhancing the Existing Infrastructure.....	2
2.1 Accounting	2
2.1.1 <i>Central Accounting Repository</i>	2
2.1.2 <i>Storage Accounting and Reporting</i>	3
2.2 PRACE RI Web and PRACE Event System Integration.....	3
2.2.1 <i>Initial Status</i>	4
2.2.2 <i>Technical Background</i>	4
2.2.3 <i>SPIP Plugins and Features</i>	5
2.2.4 <i>Implementation</i>	5
2.3 Service Certification	5
2.4 DECI Portal	6
2.5 PRACE Information Portal.....	7
2.6 INCA Monitoring	8
2.7 Collaboration with other technological oriented projects.....	9
2.7.1 <i>MAPPER</i>	9
2.7.2 <i>EMI</i>	9
2.7.3 <i>IGE and EGCF</i>	10
2.7.4 <i>EGI</i>	10
2.7.5 <i>Pilots with user-communities on data requirements</i>	10
3 Evaluating Data Services	12
3.1 Data Strategy	12
3.2 New File Transfer Technologies.....	13
3.2.1 <i>Common Methodology</i>	14
3.2.2 <i>Data transfer tools</i>	16
3.2.3 <i>Testbed definition and preliminary results</i>	16
3.2.4 <i>Outcomes and lessons learnt</i>	17
3.3 iRODS – integrated Rule Oriented Data System.....	18
3.3.1 <i>Technical evaluations</i>	18
3.3.2 <i>iRODS Workshop</i>	20
3.3.3 <i>Involment in the Data Strategy working group</i>	20
3.3.4 <i>EUDAT collaboration and pilot projects</i>	21
3.3.5 <i>Conclusions</i>	21
3.4 File System Technologies	22
3.4.1 <i>The Use-case and the Purpose of the Evaluation</i>	22

3.4.2	<i>Technical Requirements</i>	22
3.4.3	<i>Search Phase</i>	22
3.4.4	<i>Test Environment</i>	23
3.4.5	<i>Deployment and Feature Validation Testing Methodology</i>	23
3.4.6	<i>Test Results</i>	23
3.4.7	<i>Detailed Description of the Performance Measurement</i>	24
3.4.8	<i>Performance Measurement Results</i>	25
3.4.9	<i>Conclusions and Plan for Further Work</i>	27
4	Remote Visualization	27
4.1	Introduction	27
4.2	Teradici PCoIP setup at SNIC/LU	28
4.3	CINECA Remote Connection Manger	30
4.4	Performance evaluation of VNC based remote visualization services	30
5	Summary and Future Work	32
6	Annex	33
6.1	PRACE Event Integration – Screenshots	33
6.2	DECI Portal	36
6.2.1	<i>DECI peer review tools functionality comparison table</i>	36
6.2.2	<i>DECI peer review tool functionalities requirements list</i>	40
6.3	Questionnaire on Big Data	41
6.4	Methodology for File Transfer Evaluation	44
6.4.1	<i>Introduction</i>	44
6.4.2	<i>Definitions</i>	45
6.4.3	<i>Hardware and Configuration Requirements</i>	45
6.4.4	<i>Methodology</i>	47
6.4.5	<i>Test cases</i>	49
6.4.6	<i>Template for testing the data transfer tool</i>	49
6.5	Performance Measurement of Remote Visualization	50
6.6	iRODS Evaluation Forms	52
6.6.1	<i>iRODS Workflow-Objects Evaluation</i>	52
6.6.2	<i>iRODS PAM-LDAP-Authentication-Evaluation</i>	59
6.6.3	<i>iRODS-Ticket-Based-Access-Evaluation</i>	68
6.6.4	<i>iRODS FUSE-Evaluation</i>	70
6.6.5	<i>iRODS Performance Evaluation</i>	72
6.6.6	<i>iRODS - Direct Access Resources Evaluation</i>	77
6.6.7	<i>iRODS iDROP evaluation</i>	82

List of Figures

Figure 1:	Architectural diagram of the VPH use case pilot.....	11
Figure 2:	File System Performance Comparision (Raw Blocks)	26
Figure 3:	File System Performance Comparision (Ext4, Ceph, GlusterFS).....	27
Figure 4:	Images compressed with lossless zlib, lossless jpeg, and default settings.....	31
Figure 5:	Images with jpeg compression with WAN setting, 12%, and 7% setting.....	32
Figure 6:	Event Integration Screenshot 1 – PRACE Training Events.....	33
Figure 7:	Event Integration Screenshot 2 – Upcoming PATC Courses	34
Figure 8:	Event Integration Screenshot 3 – PATC Courses	35
Figure 9:	DECI-PPR-Tool Screenshot	36
Figure 10:	Graphical example for narrow and tight network links	45

List of Tables

Table 1: DECI-PPR-Tool Functionality Requirements.....	6
Table 2: Test benches for evaluating new file transfer technologies	16
Table 3: iRODS-testbed characteristics.....	18
Table 4: Matrix of test cases for file systems.....	25
Table 5: Test benches for evaluating new file transfer technologies	39
Table 6: DECI-PPR-tool complete requirement list.....	41
Table 7: File Transfer Measures Definitions.....	45
Table 8: File Transfer Requirements list.....	47
Table 9: File Transfer test cases with at least 18 runs each	49
Table 10: Example of a filled file transfer evaluation sheet, here for bbcp between CINES and CEA	50

References and Applicable Documents

- [1] PRACE project web-site: <http://www.prace-project.eu>
- [2] PRACE research infrastructure web-site: <http://www.prace-ri.eu>
- [3] PRACE-2IP deliverable D10.1: http://www.prace-ri.eu/IMG/pdf/D10-1_2ip.pdf
- [4] Grid-SAFE: <https://prace-acc.epcc.ed.ac.uk/prace/GridSAFE>
- [5] Grid-SAFE documentation:
<http://gridsafe.forge.nesc.ac.uk/Documentation/GridSafeDocumentation/>
- [6] Globus GridFTP: <http://www.globus.org/toolkit/docs/latest-stable/gridftp/>
- [7] tgftp: <http://work.deisa.eu/svn/general/globus/GridFTP/tgftp/current/README>
- [8] gtransfer: <http://www.prace-ri.eu/Data-Transfer-with-gtransfer>
- [9] GlobusOnLine: <http://www.globusonline.org>
- [10] Unicore FTP: <http://www.unicore.eu/documentation/manuals/unicore6/files/uftp/>
- [11] EUDAT – European Data Infrastructure: <http://www.eudat.eu>
- [12] PRACE-1IP deliverable D6.3 “Second Annual Report on the Technical Operation and Evolution” (2012): http://prace-ri.eu/IMG/pdf/d6.3_1ip.pdf
- [13] VPH: <http://vip.creatis.insa-lyon.fr:8080/VPH-EP-9>
- [14] FasterData Project: <http://fasterdata.es.net>
- [15] Recommended settings for TCP variables: <http://www.frozentux.net/ipsysctl-tutorial/ipsysctl-tutorial.html#TCPVARIABLES>
- [16] Enabling High-Performance Data Transfer, PSC:
<http://www.psc.edu/index.php/networking/641-tcp-tune>
- [17] BBCP: <http://www.slac.stanford.edu/~abh/bbcp/>
- [18] ARC: http://wiki.nordugrid.org/index.php/ARC_middleware
- [19] SweStore: <http://snicdocs.nsc.liu.se/wiki/SweStore>
- [20] PRACE-1IP deliverable D4.3.2 “Cross-National Programme for Tier-1 Access Pilots”
http://www.prace-ri.eu/IMG/pdf/d4.3.2_1ip.pdf
- [21] SPIP web-documentation system: <http://www.spip.net/>
- [22] InDiCo: <http://indico-software.org/>
- [23] InDiCo documentation: <http://indico.cern.ch/ihelp/html/index.html>
- [24] OGF GLUE 2.0 Specification: <http://www.ogf.org/documents/GFD.147.pdf>
- [25] REST: http://en.wikipedia.org/wiki/Representational_State_Transfer
- [26] JSR 268: http://en.wikipedia.org/wiki/Java_Portlet_Specification
- [27] INCA-Monitoring: <http://inca.sdsc.edu/>
- [28] DART: <http://www.prace-project.eu/Accounting-Report-Tool>
- [29] MAPPER: <http://www.mapper-project.eu>
- [30] IGE project: <http://www.ige-project.eu/>
- [31] Globus Toolkit: <http://www.globus.org/toolkit/>
- [32] EGCF: <http://www.egcf.eu/>

- [33] EGI (European Grid Infrastructure): <http://www.egi.eu>
- [34] EUDAT: <http://www.eudat.eu/>
- [35] iRODS-workshop: <http://www.prace-ri.eu/iRODS-workshop>
- [36] iRODS-workshop Agenda: https://www.irods.org/index.php/iRODS_User_Group_Meeting_2013#iRODS_User_Meeting_Agenda
- [37] PAM: http://en.wikipedia.org/wiki/Pluggable_Authentication_Modules
- [38] FUSE: http://en.wikipedia.org/wiki/Filesystem_in_Userspace
- [39] iRODS-FUSE module: https://www.irods.org/index.php/iRODS_FUSE
- [40] Gluster-FS: <http://www.gluster.org/docs>
- [41] Coda-FS: <http://www.coda.cs.cmu.edu/doc/html/index.html>
- [42] Ceph-FS: <http://ceph.com/docs>
- [43] Gfarm-FS: <http://datafarm.apgrid.org/document/>
- [44] Xtream-FS: http://www.xtreemfs.org/all_features.php
- [45] Lustre-FS: http://wiki.lustre.org/index.php/Lustre_Documentation
- [46] SAM-FS: <https://wikis.oracle.com/display/SAMQFS/home/>
- [47] Bug in Gfarm: <http://sourceforge.net/apps/trac/gfarm/ticket/505>
- [48] FIO: <http://freecode.com/projects/fio>
- [49] Latency: <http://www.ciinow.com/2013/01/the-truth-about-latency-in-cloud-gaming/>
- [50] VirtualGL: <http://www.virtualgl.org/>
- [51] TurboVNC: <http://www.virtualgl.org/Downloads/TurboVNC>
- [52] RCM, viz service user doc: <http://www.hpc.cineca.it/content/remote-visualization>
- [53] Teradici PCoIP technology: <http://en.wikipedia.org/wiki/PCoIP>
- [54] ParaView: <http://www.paraview.org/>
- [55] Blender modeling tool: <http://www.blender.org/>
- [56] MeshLab clout point and mesh editor: <http://meshlab.sourceforge.net/>
- [57] Visit visualization tool: <https://wci.llnl.gov/codes/visit/>
- [58] OpenCV Computer Vision Library: <http://opencv.org/>
- [59] UniGine rendering engine: <http://unigine.com/>
- [60] CD-adapco STAR-CCM+ CFD tool: <http://www.cd-adapco.com/products/star-ccm-plus>

List of Acronyms and Abbreviations

AAA	Authorization, Authentication, Accounting.
ADSL	Asynchronous Digital Subscriber Line
AISBL	Association International Sans But Lucratif (legal form of the PRACE-RI)
AMD	Advanced Micro Devices
API	Application Programming Interface
ARC	Advanced Resource Connector
BDP	Bandwidth Delay Product
BSC	Barcelona Supercomputing Center (Spain)
CEA	Commissariat à l'Énergie Atomique (represented in PRACE by GENCI, France)
CINECA	Consorzio Interuniversitario, the largest Italian computing centre (Italy)
CINES	Centre Informatique National de l'Enseignement Supérieur (represented in PRACE by GENCI, France)
CNRS	Centre National de la Recherche Scientifique
CPU	Central Processing Unit
DANTE	Delivery of Advanced Network Technology to Europe
DART	Distributed Accounting Record Tool

DCV	Deep Computing Visualization (IBM) or Desktop Cloud Virtualization (NICE)
DECI	Distributed Extreme Computing Initiative
DEISA	Distributed European Infrastructure for Supercomputing Applications. EU project by leading national HPC centres.
DoE	United States Department of Energy
DPMDB	DECI Project Management Database
EGCF	European Globus Community Forum
EMI	European Middleware Initiative
EPCC	Edinburg Parallel Computing Centre (represented in PRACE by EPSRC, United Kingdom)
EPSRC	The Engineering and Physical Sciences Research Council (United Kingdom)
FIO	Flexible I/O Tester
FUSE	File System in User Space
FZJ	Forschungszentrum Jülich (Germany)
GB	Giga (= $2^{30} \sim 10^9$) Bytes (= 8 bits), also GByte
Gb/s	Giga (= 10^9) bits per second, also Gbit/s
GB/s	Giga (= 10^9) Bytes (= 8 bits) per second, also GByte/s
GCS	Gauss Centre for Supercomputing (Germany)
GÉANT	Collaboration between National Research and Education Networks to build a multi-gigabit pan-European network, managed by DANTE. GÉANT2 is the follow-up as of 2004.
GENCI	Grand Equipement National de Calcul Intensif (France)
GFlop/s	Giga (= 10^9) Floating point operations (usually in 64-bit, i.e. DP) per second, also GF/s
GHz	Giga (= 10^9) Hertz, frequency = 10^9 periods or clock cycles per second
GigE	Gigabit Ethernet, also GbE
GNU	GNU's not Unix, a free OS
GPU	Graphic Processing Unit
GSI	Grid Security Infrastructure
GSI-SSH	GSI based ssh (secure shell, remote login facility)
HPC	High Performance Computing; Computing at a high performance level at any given time; often used synonym with Supercomputing
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
InDiCo	Integrated Digital Conference
I/O	Input/Output
IPB	Institute of Physics Belgrade
iRODS	Integrated Rule-Oriented Data System
JSC	Jülich Supercomputing Centre (FZJ, Germany)
KB	Kilo (= $2^{10} \sim 10^3$) Bytes (= 8 bits), also KByte
LLNL	Lawrence Livermore National Laboratory, Livermore, California (USA)
LRZ	Leibniz Supercomputing Centre (Garching, Germany)
MAPPER	Multiscale Applications on European e-Infrastructures
MB	Mega (= $2^{20} \sim 10^6$) Bytes (= 8 bits), also MByte
MB/s	Mega (= 10^6) Bytes (= 8 bits) per second, also MByte/s
MFlop/s	Mega (= 10^6) Floating point operations (usually in 64-bit, i.e. DP) per second, also MF/s

MHz	Mega (= 10 ⁶) Hertz, frequency =10 ⁶ periods or clock cycles per second
Mop/s	Mega (= 10 ⁶) operations per second (usually integer or logic operations)
MoU	Memorandum of Understanding
MPG	Max-Planck-Gesellschaft (Germany)
MPI	Message Passing Interface
MTU	Maximum Transmission Unit
NFS	Network File System
NIC	Network Interface Controller
NIIF	Nemzeti Információs Infrastruktúra Fejlesztési Intézet (National Information Infrastructure Development Institute, Hungary)
NSC	National Supercomputing Centre in Linköping, Sweden
OpenGL	Open Graphic Library
OS	Operating System
PAM	Pluggable Authentication Modules
PCoIP	Pixel Compression over Internet Protocol
PCIe	Peripheral Component Interconnect express, also PCI-Express
PHP	originally: Personal Home Page; now: Hypertext Preprocessor
PID	Persistent IDentifier
pNFS	Parallel Network File System
POSIX	Portable OS Interface for Unix
PRACE	Partnership for Advanced Computing in Europe; Project Acronym
PSNC	Poznan Supercomputing and Networking Centre (Poland)
RAID	Redundant Array of Independent Disks
REST	REpresentational State Transfer
RTT	Round Trip Time
RZG	Rechenzentrum Garching der MPG (Germany)
SAN	Storage Area Network
SAS	Serial Attached SCSI
SATA	Serial Advanced Technology Attachment (bus)
SGI	Silicon Graphics, Inc.
SNIC	Swedish National Infrastructure for Computing (Sweden)
SNIC/LiU	Swedish National Infrastructure for Computing/Linköping University
SPIP	Système de Publication pour l'Internet Partagé
SSD	Solid State Disk or Drive
STS	Security Token Service
SURFsara	Dutch national High Performance Computing & e-Science Support Center
TB	Tera (= 240 ~ 1012) Bytes (= 8 bits), also TByte
TFlop/s	Tera (= 1012) Floating-point operations (usually in 64-bit, i.e. DP) per second, also TF/s
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
UNICORE	Uniform Interface to Computing Resources. Grid software for seamless access to distributed resources.
USB	Universal Serial Bus
VIP4VPH	Virtual Imaging Platform for the Virtual Physiological Human
VM	Virtual Machine
VNC	Virtual Network Computing
XML	eXtended Markup Language

Executive Summary

The major aim of Work Package 10 (WP10) *Advancing the Operational Infrastructure* in PRACE-2IP is assessing software technologies and promoting services needed for the operation of the integrated PRACE-infrastructure. WP10 partly continued work of the technology task T6.3 of work package WP6 of PRACE-1IP. The results will be handed over to T6.3 *Technical evolution of the PRACE services* of WP6 *Operation of the Distributed Infrastructure* in PRACE-3IP.

This work is organised in three tasks covering the three goals to be achieved. The first task focused on enhancing the existing infrastructure. The second one concentrated on data services, which go beyond the already existing ones. The third task covered the remote visualization of data.

In the first task, the following services have been treated in several sub-tasks: The centralized accounting service *Grid-SAFE* has been extended covering now Tier-0 and Tier-1 systems and will go into production soon. Another sub-task integrated the announcement and management of PRACE events directly into the PRACE web-site. Work on the *Service Certification*, the *PRACE Information Portal* and the *INCA Monitoring* has been continued by further developing the products or new or additional components for them and thus led to an overall improved infrastructure. The Project Proposal Revision (PPR) tool, already in use for managing Tier-0 calls, has been extended to also handle proposals of the DECI calls in Tier-1. It has already been used for the latest DECI-11 call. Finally, the collaboration with other technologically oriented projects has been continued, especially with concrete pilot projects addressing data requirements of user communities. One pilot dealt with the needs of the Virtual Imaging Platform for the Virtual Physiological Human (VIP4VPH) [13], the other one handled the complex data-workflow of a quantum mechanical problem using local, EGI, and PRACE-resources for the calculation as part of the MAPPER-project [29].

In the second task, *Data Services*, the *Data Strategy* group – as one sub-task – generated a questionnaire on Big Data and proposed different recommendations for improvements in data handling, of which some could be implemented easily. The sub-task *New File Transfer Technologies* compared several file-transfer-tools, currently not yet supported by PRACE in the service-catalog, and assessed their potential benefits for users in PRACE. Initially the iRODS repository technology had been considered the most promising software concerning handling of data identified by persistent identifiers and was therefore included into the software to be evaluated by WP10. As the pilot collaboration with EUDAT showed, there is a real user need for such a technology. Therefore it has been evaluated very deeply. Depending on further user or community requests iRODS may become a generally supported service, at least optional, in PRACE. Finally, several file system technologies have been investigated with respect to their possible use in PRACE. The performance measured and the stability tested does not yet allow recommendations of any of the evaluated file system technologies.

The third task, *Remote Visualization*, continued its investigations in different technological implementations based on VNC technologies. The Remote Connection Manager pilot installation has been finalized and will become a production service in PRACE. Furthermore, performance has been tested for varying hardware and network configurations with different software implementations being able to give recommendations on the best setup and usage of VNC based remote virtualization depending on the infrastructural conditions.

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 organised in sub-tasks to better focus the specific topic.

Structure of the Document

The following document consists of three further chapters – *Enhancing the Existing Infrastructure*, *Evaluating Data Services*, and *Remote Visualization* – 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. A chapter *Summary and Future Work* will conclude, and finally an *Appendix* with several sections provides 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. As in PRACE-1IP again in PRACE-3IP the technological evolution is covered as task T6.3 of WP6, while in PRACE-2IP the separate work-package WP10 was dealing with technological developments. The deep collaboration between WP10 and T6.3 is achieved by having joint bi-weekly video-conferences coordinating the work.

2 Enhancing the Existing Infrastructure

The objective of task 10.1 is to identify and evaluate options for technical enhancements to the existing Tier-1 services. Input did come from within the work package, other work packages, like WP2 for the *DECI Portal* (see 2.4) for the handling of the DECI proposals, or as a result of surveys, as for the *Storage Accounting and Reporting* (see 2.1.2). Furthermore, through the collaborations direct user or user community requests, as the pilots together with EGI and EUDAT (see 2.7.5), influenced the working directions of this task.

2.1 Accounting

Current accounting covers CPU-usage only. In this area improvements for the storing of the accounting information and the easy access to it for users are a major task. In addition, with the increasing amount of data produced more and more considerations come up to also include accounting of storage usage.

2.1.1 Central Accounting Repository

A centralized accounting service was set up in previous years using the *Grid-SAFE* tools developed by EPCC [4]. In October 2012 a document was produced as input for the acceptance procedure as a production service. Based on this input the members of the operation groups of all sites, both Tier-0 and Tier-1, have accepted in November 2012 to propose to PRACE management the *Grid-SAFE* facility as a production service with classification *additional* as defined by the PRACE Service Catalogue. The service is included

in a new version 2.4 of the Service Catalogue which is submitted for acceptance to the PRACE Hosting Members by WP6 of PRACE-3IP.

In October 2012 the PRACE Security Forum completed a risk review of the new service with as result that there was no objection to run this service.

User documentation also was produced and reviewed and will be published once the service is going into full production.

Pre-production tests have been prepared and run. The results have been used to correct errors for some sites.

An additional feature was added, which enables partners to start automatically a new update for the last months. This can be needed if local data was updated, e.g. because usage was reimbursed for jobs.

The development of a certification procedure for *Grid-SAFE* was started in the sub-task *Service Certification* but has to be completed yet.

All partners that have a local PRACE accounting service can now export data to the central service. At the moment 14 partners export their data.

2.1.2 Storage Accounting and Reporting

The objective of this activity was to analyse the need and possibilities of storage accounting and disk usage information for users.

A survey was prepared in the first project period to collect information from all PRACE partners and AISBL on this subject. The survey (see appendix in deliverable D10.1 [3]) was issued to all partners/sites in October 2012, of which twenty partners have responded. The results have been processed and this resulted in a report by the end of 2012. The internal report and its conclusions have been discussed in two dedicated video conferences early in the spring of 2013.

The report gives an overview of the disk storage accounting policies and tools in use by sites. The main conclusions and results of the report are:

- Less than half of the partners use disk storage accounting.
- The tools that are provided to users to get information on available and used storage vary from built-in OS tools and specific file system tools to specific site developed scripts and open source tools.
- On the PRACE internal wiki a table is maintained with up-to-date information about the storage facilities at sites.
- As a next step it is proposed to investigate in the use of a uniform tool within PRACE for the provisioning of information about actual storage utilization to users.
- There is no requirement from partners to further develop storage accounting facilities.

Task 6.3 of PRACE-3IP-WP6 can use the results of this activity to further evaluate and develop facilities to display the actual storage utilization.

2.2 PRACE RI Web and PRACE Event System Integration

The integration of an event-managing-system into the regular PRACE-RI Web-Site came as a response to the ever increasing need to announce the PRACE events in a more efficient manner and to enhance ease of access to all the relevant information. Integrating these services required development related activities which were successfully performed in WP10.

2.2.1 *Initial Status*

PRACE RI Web-Site

The PRACE RI main website is based on SPIP CMS [21]. It is hosted at and administered by CINES. The software has a GPL license and documentation is mostly in French. It is written in PHP with a MySQL database. New features can be added as plugins through a website backend (available to web administrators) or by storing it directly on the machine (available only to CINES staff).

PRACE Events System

PRACE Events System is based on InDiCo software [22]. InDiCo is a web application for scheduling and organizing events, from simple lectures to complex meetings, workshops and conferences with various sessions and contributions. It was originally developed in the framework of the EU InDiCo project [22], but currently, InDiCo is free software licensed under terms of GNU General Public License (GPL). The InDiCo user guide can be found at [23]. The PRACE Events System is hosted at and administered by IPB.

2.2.2 *Technical Background*

Integration

The aim is to enable automatic display on the PRACE RI website of upcoming events that are entered in the InDiCo System. Events should be displayed in three categories:

- Upcoming events, sorted by date, first to come is on top
- Past events, sorted by date, latest on top
- Calendar view, all events sorted by year, month day or in a calendar view

Exporting Data

InDiCo provides several ways to export data:

- To Personal Scheduler Tools (Outlook, iCal...)
- RSS feeds
- Sharepoint
- HTTP Export API

HTTP Export API

InDiCo allows for programmatically access to the content of its database by exposing information like category contents, events, rooms and room bookings through a web service, through the HTTP Export API. The basic URL looks like:

```
http://my.indico.server/export/WHAT/[LOC/]ID.ID.TYPE?PARAMS&ak=KEY&timestamp=TS&signature=SIG
```

where:

- WHAT is the element to export (one of categ, event, room, reservation)
- LOC is the location of the element(s) specified by ID and only used for certain elements
- ID is the ID of the element to export (can be a - separated list)
- TYPE is the output format (one of json, jsonp, xml, html, ics, atom, bin)
- PARAMS are various parameters affecting (filtering, sorting, ...) the result list
- KEY, TS, SIG are part of the API Authentication.

Details about HTTP Export API URL parameters can be found in the user manual.

2.2.3 SPIP Plugins and Features

Syndication (CMS built in feature)

The syndication system allows sharing the attached document urls (podcasting), transferring keywords (tags) from one site to the other as well as transferring the section (or category) of the articles. The default templates provided by SPIP include a RSS feed template.

Fullcalendar (plugin)

The Fullcalendar plugin creates calendars from the articles, the SPIP database or the Google calendar when included in the articles or columns.

2.2.4 Implementation

All three integration requirements (upcoming and past events and a calendar view) have been successfully met and the implemented features can be accessed and used on the prace-ri.eu website (see e.g. <http://www.prace-ri.eu/PRACE-Training-Events>).

PRACE RI website and PRACE Events System integration offers ease of access to users and direct links to PRACE Events website (<http://events.prace-ri.eu/>) for the desired events, both through events list and calendar view.

2.3 Service Certification

The main goal of the *Service Certification* sub-task was to define and implement procedures for ensuring adequate level of quality of services within PRACE infrastructure before enabling them for users. This includes verification of deployed services before offering them to the users, ensuring that technical requirements are satisfied, ensuring that quality standards, such as operational policy are satisfied, and improving the quality of offered services.

Within the reporting period the activity focused on finalizing the general certification procedure, implementing quality checklists and test scripts for selected services, and performing the certification on selected services. Currently the list of services, which have at least partial quality checklists, includes:

- Uniform access to HPC (partial)
- PRACE internal interactive command-line access to HPC (complete)
- Data transfer, storage and sharing (complete)
- Authentication (partial)
- Authorization (partial)
- Accounting (partial)
- Grid-SAFE Accounting repository (complete)
- Network management (partial)
- Monitoring (partial)
- Software Management and Common Production Environment (complete)

The main implementation details behind *Service Certification* have been decided. First of all, certification results should be stored on the wiki using special templates called the Service Certification Log. The quality checklists will be continuously improved by the sites assigned to perform the specific certifications. Finally, it was decided that certification results will be internal for PRACE staff only.

The activity will be continued in Task 6.3 of PRACE 3IP, and will include performing and improving the certification procedures and integration of selected service certification tests with INCA for procedure automation.

2.4 DECI Portal

The objective of this activity was to setup a tool to improve the submission and management of the project proposals in the DECI calls [20]. A steering committee was set up and comparing the PRACE Tier-0 PPR (developed by CINES) against the HPC Europa PPR (developed by CINECA). (See appendix 6.2.1 for the functionalities comparison table).

It was decided to implement the DECI Project Proposal Revision on a re-implemented version of the same software on which the PRACE Tier-0 peer review tool is based. Thus the same basic software is used for Tier-0 and Tier-1 calls in PRACE. Concrete work started from December 2012 where a steering committee was set-up for driving endeavours. The new portal was then first used for running the 11th DECI call for proposals.

Out of the initial functionality requirements list (see appendix 6.2.2), the following items have been addressed:

#	Functionality	Rate
1	Electronic submission of project proposals.	Essential
2	Developers' ability to programmatically redesign the forms contents and their integration with the internal database.	Essential
4	Provide users with complete online control of their data (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).	Essential
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
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)	Desiderata
13	Keep extensive logs regarding all changes made by the users in the tool.	Desiderata
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).	Desiderata
19	Create a report of all persons involved in past and present calls (PIs, collaborators) with history (call, proposal ID, ...)	Essential
20	Guarantee a highly secure log-in system (highly secure password)	Essential

Table 1: DECI-PPR-Tool Functionality Requirements

Actual adaptation work from the Tier-0 tool started on April 6th 2013 in close contact with WP2's DECI program manager for the DECI-11 call, which was open from May 6th to June 14th.

At the time this document is being edited, the tool was in use for three months. The adaptation for DECI is considered to be at an early stage, and the tool is still under evaluation together with WP2. Because of its relevance it will be continued in the technology task T6.3 of PRACE-3IP.

As of July 2013, the tool provides an implementation of a DECI peer review process with complete on-line handling of proposals from the submission to the technical and scientific review assessment.

- Electronic submission of HPC project proposals:
 - Incremental edition of proposals while a call is open
 - Declaration of project investigators (collaborators)
 - Final submission with prior checks for required fields being filled in
- Follow-up of the peer review process of a given call is mainly achieved from a «master spreadsheet» page where relevant data is display on the status of proposals and reviews
- Technical evaluation (TE) where all relevant data from proposals are visible in TE form
- Scientific Evaluation (SE) where evaluators can get limited access to relevant proposals and TE
- Having on-line forms allows for leveraging typical database abilities in terms of data extraction, such as obtaining a list of proposals that match criterion or a list of registered users
- Transverse functionalities notably encompass data export to Excel spreadsheets and PDF

Interfaces to other PRACE services will be going to be implemented step by step into the PPR-tool as part of T6.3 of PRACE-3IP:

- Interfacing with the DPMDB tool for project follow-up
- Interfacing with PRACE's central LDAP: Notably for “external” authentication of DECI staff users
- Interfacing with the Grid-SAFE/DART accounting infrastructure for follow-up of awarded projects resources consumption.

Some further DECI specific processes might need some additional supporting functionality in the tool, such as the DECI Access and Allocation Committee (DAAC) where awarded proposals are assigned to sites and actual computer systems.

2.5 PRACE Information Portal

PRACE users require various information to efficiently use services offered in the e-Infrastructure. This information includes network status and performance, HPC resource maintenance schedule, service availability and functionality and so on. PRACE operates a number of tools, e.g. *Iperf* for network and *Inca* for service monitoring, to measure and collect detailed statistics on availability and functionality of production services. This data is, to a large extent, available solely to PRACE staff members as raw data sets often contain security or privacy relevant information. This limits the sharing and distribution of e-Infrastructure state data among end-users. Annual PRACE user surveys emphasize these limitations and provide details on the kind of data users need to efficiently work in PRACE.

A sub-task of WP10 is responsible for the evaluation, design and implementation of a portal, called the *PRACE Information 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. The portal team comprises members of three PRACE sites: BSC (Spain), LRZ (Germany) and PSNC (Poland). The team is led by LRZ.

The task started with an evaluation of existing and operational technologies that could be used for providing PRACE end-users with required information. Throughout the evaluation process no suitable existing service could be identified, which led to the conclusion that the information portal has to be developed based on a novel technology.

The architecture of the information portal is largely based on standards to ensure easy integration with PRACE operational services. Some of the standards adopted or implemented in the portal include OGF GLUE2.0 [24], REST [25], and JSR268 [26].

For the first release of the portal the following five high priority requirements were chosen for implementation:

- PRACE Link Availability
- Internet Link Availability
- Core Services Status
- Account Usage
- System Information

At this moment the first implementation phase is nearing its completion. Implementation of necessary information providers is finished, logic and web interface are currently under development. PSNC has designed and implemented the following information providers:

- Network Monitoring
- INCA
- LDAP

The implementation covered the mechanisms for gathering of the information originating from different PRACE sources, processing and finally putting them into the database.

For the better understanding of the collected data and the debugging process, PSNC created a simple test web portal. It is accessible for PRACE users with their certificate under the following address: <https://dmon-prace-fe.srv.lrz.de>

More information providers will be combined into the future production PIP portal. In the test portal one can view currently:

- PRACE Link Availability
- PRACE Link Latency
- System Information (currently software versions and service availability)

Work, achievements and details on all ongoing activities of the task are thoroughly documented in PRACE Wiki.

2.6 INCA Monitoring

Within PRACE the *Inca Monitoring* – based on the *Inca* software originally developed by SDSC [27] – is constantly updated by WP6 to match the current state of the PRACE infrastructure. Besides this, within WP10 work is conducted to further improve the user interface offered by Inca monitoring. A complete coverage of all services and tools employed within PRACE should be achieved as the final goal.

Therefore, existing Inca reporters originating from the DEISA project needed to be adapted to the PRACE infrastructure. Namely the Inca reporter for the version of the FFTW library was re-implemented, tested and successfully deployed to match the new conditions.

Furthermore, to cover the complete production environment of PRACE, several new Inca reporters for different middleware tools were developed. These include the version tests for the `prace_service` script and configuration which are an essential part of the middleware services. Further Inca reporters testing for the existence and version number of the *gtransfer* tool, the *myproxy* client, the *GSI-SSH* client and the *GridFTP* client have been developed. They are either based on existing reporters or have been developed from scratch. All mentioned Inca reporters have been tested and deployed successfully.

In addition, a new Inca reporter prototype for the PRACE accounting infrastructure based on *DART* [28] was developed. It is currently in testing stage and will be transferred into production soon. It is currently evaluated if it may serve as a template for monitoring the *Grid-SAFE* based accounting as well.

2.7 Collaboration with other technological oriented projects

Since the beginning of the project, PRACE has actively collaborating with other e-Infrastructures and EU project to improve users experience, strengthen the collaboration with external technology providers, exchange knowledge among technical people, raise the awareness around PRACE services and disseminate its activities. Over the course of the second year of the project a new collaboration was initiated with the EUDAT project also involving the EGI infrastructure and new scientific communities resulting in a few pilots. The following subsections give an overview of the collaborations of which some will continue within the Task 6.3 of PRACE-3IP.

2.7.1 MAPPER

The MAPPER project (Multiscale APplications on EuROpean e-infrastructures) [29] 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 started in May 2011 and was coordinated via a Task Force comprising specialists from each of the three organisations (MAPPER, PRACE, EGI-Inspire).

On request of MAPPER then PRACE and EGI investigated the exchange of user support requests between the EGI and PRACE helpdesks. This should provide end users the ability to request support from both infrastructures with just one request. Technical requirements to enable such exchange of support requests between the two helpdesks have been exchanged between the two projects but the implementation has still to be planned.

Some of the user communities involved in the MAPPER project applied for preparatory access to PRACE facilities. These allocations give these communities the opportunity to submit jobs through MAPPER developed tools. The research that the MAPPER project is pursuing has a distributed nature, binding different communities and systems. Access to the PRACE e-Infrastructure is required to ensure proper functionality and integration of tools and services provided by MAPPER. In particular, the focus is on the software/middleware development and adaptation, taking care of the infrastructure management rather than resource provisioning.

2.7.2 EMI

The EMI (European Middleware Initiative) 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 projects, and to extend the interoperability and integration between grids and other computing

infrastructures. A joint work-plan to implement collaboration's objectives was defined in a Memorandum of Understanding (MoU) which was officially signed by respective projects coordinator at the beginning of 2013. The EMI officially ended on April 2013. As part of the work-plan, EMI components, those belonging to the UNICORE platform, have been officially tested on PRACE sites (CINECA, FZJ) and feedback, in form of requirements, have been sent to EMI STS (Security Token Service) developers.

2.7.3 IGE and EGCF

The Initiative for Globus in Europe (IGE) [30] did support 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* [31], including *GridFTP* and *GSI-SSH* which are currently deployed in PRACE. A MoU, which describes the activities of the collaboration, was signed early this year. For the evaluation of the *GlobusOnLine* tool IGE provided feedback on questions and problems. Another important activity is the support for the Globus tools in production by PRACE. The IGE project ended March 2013; however IGE support activities are continued by the European Globus Community Forum (EGCF) [32].

2.7.4 EGI

With the European Grid Infrastructure (EGI) [33], besides the interoperation of the helpdesks, also the exchange of resource usage information was discussed. This will provide user communities that use resources in different infrastructures a single view of their resource usage and can help users in choosing the most appropriate resource to run their jobs. The technical details to enable such an exchange have been discussed between PRACE and EGI; however the implementation is waiting on efforts by EGI.

2.7.5 Pilots with user-communities on data requirements

A new activity named *Data sharing and uniform data access across e-infrastructures and community centres* between PRACE, EGI and EUDAT [34] started this period to address issues of data management interoperability. The objective was to identify use cases of user communities that need to share data among these three infrastructures and to identify limitations and requirements using these use cases. In November 2012, together with EGI and EUDAT a two day workshop was organised in Amsterdam where several user communities with interoperability needs (e.g. VPH, EPOS, ENES, MAPPER, ScalaLife, VERCE, DRIHM, MSS) have been invited to present their use cases. The infrastructures in return presented their data management facilities and plans. As a result of this workshop a few pilot projects have been defined, each with a specific objective and with the involvement of one or more user communities and representatives of the infrastructures.

They all aimed at establishing a prototype to share data across the e-Infrastructures and community centers for medium term storage. An integration workflow driving the pilot activity is typically composed by the following steps:

- data sets are ingested and registered onto EUDAT resources; this will assign a persistent identifier (PID) to the data;
- data identified by this PID are then staged onto computational resources for further processing. PRACE resources are usually utilized for massive data processing while EGI ones for post-processing;
- results produced are ingested back on EUDAT and included in community data collections.

At the moment, two use cases with PRACE involvement are being implemented following VPH and MAPPER requirements. Only mature communities have been effectively involved into the pilots for basically two reasons: a) to limit the effort into few months of work so to only focus on achievable and operative goals, b) to work with communities which already experimented the integration of different services and got stucked really close to complete their plan.

VIP4VPH

The goal of this project is to offer imaging scientists a convenient mechanism to access computational and data resources ensuring the sustainability of image simulation workflows beyond a particular computing infrastructure and workflow technology. This is realized providing an interoperability layer between the Virtual Imaging Platform and the VPH toolkit [13]. Multi-modality medical image simulators (MRI, US, CT, and PET) are described as workflows using the MOTEUR technology which is able to access several infrastructure services seamlessly. The pilot worked to achieve two main goals: a) identify which sites, being part of involved infrastructures, could commit their resources for the community; b) develop a mechanism to easily ship data across the sites. To address the second goal the GridFTP protocol was selected and client adapted to handle data transfer using the EUDAT PID. Currently, the participating sites have been confirmed (EPCC will make available its PRACE resources) and data transfer performance across the sites are under evaluation.

MAPPER

The goal of the project is to develop computational strategies, software and services for distributed multiscale simulations across disciplines, exploiting existing and evolving European e-Infrastructure. The diagram below presents the steps of a typical simulation workflow and the resources potentially involved and belonging to different e-Infrastructures. The diagram was elaborated during the pilot activity.

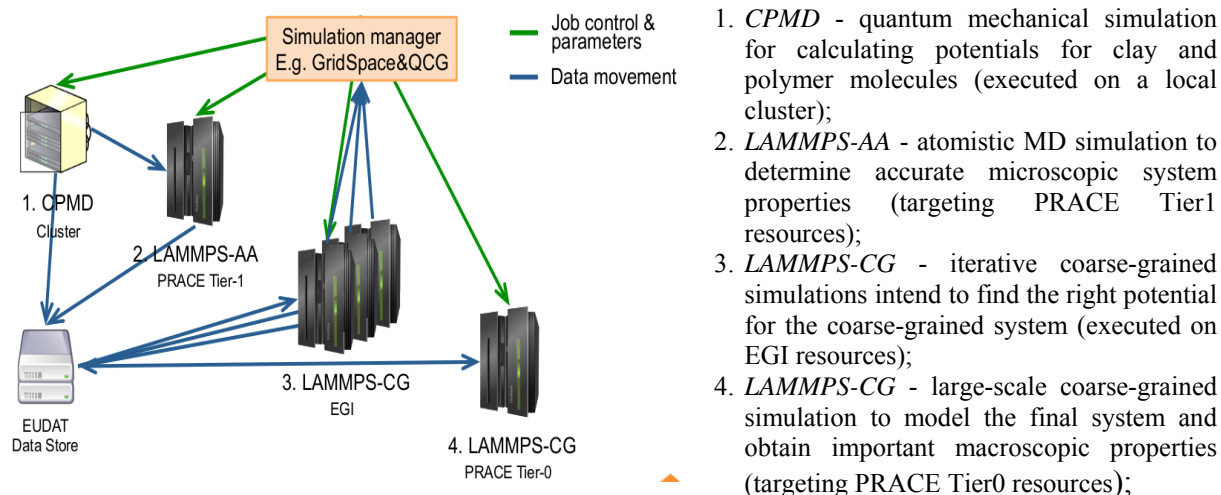


Figure 1: Architectural diagram of the VPH use case pilot

A collaboration with this project was already in place but after the EEP (EUDAT-EGI-PRACE) workshop it was reorganized and merged under this broader collaboration umbrella. The only site currently involved for PRACE is LRZ which hosts the software components (QCG-Computing, MUSCLE, QCG-Broker, QCG-Notification) necessary to execute MAPPER workflows.

Conclusions

In March 2013 a one day workshop was organised in London for a first evaluation of the progress and to plan the next steps. For PRACE the most important results of the pilots are:

- *GridFTP* is the data transfer protocol that can be used on all three infrastructures;
- The use of *Persistent Identifiers* (PID) for the registration and access of data from PRACE on the EUDAT infrastructure was implemented.

The pilots have formally ended August 2013, but the collaboration among the three infrastructures will continue when needed.

3 Evaluating Data Services

3.1 Data Strategy

Initial Situation

The ever increasing amount of data produced in projects computed on PRACE resources makes it more and more difficult to transfer the data in and out of the PRACE systems. The increase of the data volume is growing drastically faster than the available data transfer speed. Therefore, a more flexible data strategy in PRACE is very important.

Reasons for the need for a Data-Strategy in PRACE

- PRACE users spend more and more time on storage issues.
- Volume grows rapidly.
- Moving data between storage systems can no longer be handled by users interactively.
- Open data access demands.
- No uniform handling of data inside PRACE
- Including other partners handling data: EUDAT, National storage, etc.
- Central handling of distributed storage can better utilize available resources.
- PRACE users should focus on scientific work not data-management.

Compiling a Questionnaire

In order to better identify the real needs of the users this task generated – on request of the PRACE Management Board – a questionnaire to cover all the data related issues identified by the users of the PRACE infrastructure. This questionnaire, compiled in March and April 2013 by data-experts from several PRACE-sites, can be found in the appendix 6.3. It was accepted by the Management Board, but there was not yet a decision on when and how to perform it. Thus no answers to evaluate and base recommendations on are available yet.

But nevertheless during the discussions when developing the questionnaire already some possible recommendations, relatively easy to implement, but with a possible large impact for the users, could be identified.

Presumably Straight Forward Improvements

A uniform way of handling data inside PRACE would be beneficial. This could also include the creation of project storage with bigger volume for sharing data in one site between PRACE users. The timed availability for preserving the data within that project storage could be more flexible, e.g. allowing for storing data between different project periods (DECI-calls).

Separate data transfer-nodes as front nodes to clusters could reduce the load on the login nodes, speed up the data-transfers and eventually also allow for some sort of background data transfers.

Collaboration with providers of other data storage, like national storage systems, community storage and project storage, and providing fast connectivity to them could improve transfer speeds for data movements in and out of PRACE internal storage systems.

Finally, PRACE support could provide users with more specific help for individual needs concerning data-transfers to and from PRACE systems.

The realizations of such options require respective decisions on the executive level of PRACE followed by the implementation on the operational level.

Other – more long-term – Options

More effort is required for some of the other possible options, which sometimes require even changes in the policies and service provisioning models.

- Access to analysis and visualization nodes closely coupled with the HPC systems: The need for data movement can be reduced if not avoided and the workload for the users decreases dramatically for some projects (c.f. Remote Visualization in 4).
- Availability of intermediate preservation of data for volumes and sharing: With the provisioning of intermediate storage, data can be shared more easily between sites and big volumes can be handled cheaper. The intermediate storage can be handled in a centralized way with decentralized resources (cf. File System Technologies in 3.2).
- Provision of technologies like iRODS (cf. 3.3) also for intermediate storage.

Conclusions

It is evident that PRACE requires a clear strategy towards the handling of data. This was acknowledged on the management level, which requested the generation of the questionnaire. Further investigations and effort into the development of a profound basis for decisions are dependent on the results of the questionnaire. This needs to be distributed to the users and communities and then the collected results have to be summarized properly.

3.2 New File Transfer Technologies

The objective of this sub-task has been to carry out technical evaluations of high-performance file transfer tools in order to identify possible and reliable alternatives to Globus *GridFTP* [6], which is the only trusted and supported tool in PRACE for moving large amounts of data.

Continuous growth in computing power is increasing the need of having a reliable data transfer service for transferring bulk data inbound and outbound of the PRACE Research Infrastructure. This requirement is especially pressing for scientific data-intensive applications like those belonging to earth and life sciences.

In the past the focus was mainly on improving performance, usage and logging features of *GridFTP* [7]. Then the tool *gtransfer* [8] built on top of *GridFTP* was developed for moving data with optimized performance by an easy to use interface. Feedback received from internal surveys from DECI users and from user communities like those participating in the EUDAT project [11] pointed out that more sophisticated tools for high-performance data transfers are needed. Tests results for *GlobusOnLine* [9] and *UnicoreFTP* [10] are already documented in the PRACE-1IP deliverable D6.3 *Second annual report on the technical operation and evolution* [12]. This activity started by taking into account all these previous experiences.

The main strategy for this subtask in WP10 has been extending a perspective traditionally centered on a specific tool and considering all variables that can have an impact on transferring bulk data. This allowed defining a common methodology for carrying out tests that can be applied to future and further tests of new software solutions.

The methodology considers factors as:

- **Type of Dataset**, because moving many small files is different from moving single large files, from the perspective of I/O operations required;
- **Type of Workload**, because tools usually have different behaviours with different size of data to be transferred;
- **Host configuration**, for defining and setting up a minimum set of technical requirements for hosts involved in the communication in order to mitigate, or eliminate, bottleneck effects;
- **Network capacity**, it is generally difficult to analyse an entire network path connecting sender and receiver, but some measures along with network diagnostic tools as well as a deep understanding of the network topology can help discover the presence of any narrow-link in the middle.

After defining a common methodology, next steps have been the selection of a set of tools and the set-up of test benches.

The activity has performed already several tests, but it is too early to provide a summary and conclusion about the results, this will be done after all tests have been finalized in task 6.3 of PRACE-3IP. Then a separate document will be produced. The complete current description of the methodology, including all its definitions, can be found in the appendix 6.4, while the next section cites relevant parts of the appendix 6.4.4 describing the methodology.

3.2.1 *Common Methodology*

A common methodology for evaluating new file transfer technologies (where “new” stands for “not officially supported in PRACE”) has been designed to be independent from a specific software solution. Similar work carried out in other scientific contexts has been also taken into account to ensure robustness and completeness in the way of making tests [14]. The following factors have been considered as essential features of the methodology:

- Assessments must be produced in a consistent manner across different sites and different network paths;
- Production conditions and any network turbulence must be considered;
- Performance must be measured with different types of workloads and different numbers of parallel streams (only tools supporting parallel data streams must be considered);
- A template must be available for collecting results of tests;
- A mechanism for qualifying and comparing results must be available;
- Each test must follow specific and defined test-cases;

In addition to a quantitative assessment, factors like reliability, footprint or intrusiveness, maintenance, code maturity, support responsiveness, have been considered important as well. Both the PRACE network and the public internet have been considered as target medium for testing.

Bandwidth Delay Product

The Bandwidth Delay Product (BDP) has been selected as the measure to reflect a production condition. BDP is a term primarily used in conjunction with the TCP to refer to the number of bytes necessary to fill a TCP "path", i.e. it is equal to the maximum number of simultaneous bits in transit between the transmitter and the receiver. The BDP formula includes network capacity and round trip time (RTT) of TCP packets according to the formula:

$$\text{BDP (GByte)} = \text{Capacity (Gbps)} * \text{RTT (s)} / 8$$

It gives a measure of the network congestion, at a specific time, and the ability to compare different file transfer tools under similar values for the BDP. It was a must to calculate it before running a test.

TCP Tuning

Configuring TCP parameters for data transfer hosts is probably both the most important and the most complicated action for improving performance in the same time. Settings must take into account the available network bandwidth. But too specific configurations can sometimes even lead to a depletion of performance especially if low-speed networks are used.

Modern operating systems meanwhile provide an excellent auto-tuning for the TCP buffers leaving a system administrator play with maximum values only. Many other TCP-related variables could be recommended, but this is out of scope and similar works are suggested in [14], [15], and [16]. This activity focused on three important settings (details can be found in appendix 6.4.3):

- **TCP Buffer Size:** Values varied depending on the available memory of the machine and the BDP (see above).
- **MTU and Jumbo Ethernet Frames:** Testing with a MTU of 9000 compared to the default MTU of 1500.
- **Disk performance:** Check performance of the disks subsystem with I/O benchmarks like *hdparm*, *bonnie++* and *iozone*.

Data sets

Transferring a large number of small files is significantly different from transferring few large files in terms of performance. Also the directory tree affects performance significantly. The methodology considers two different datasets, one with many small files and the other with a few large files. Details can be found in the appendix 6.4.4.

Workload

Testing a tool against different workloads is a good way for producing an exhaustive assessment, since it simulates a wide variety of situations happening on real systems and therefore allows for detailed analysis of the behaviour in real life. A huge workload can thus provide information about stability and reliability of a software solution as well as features like checkpoint and restart of a file transfer after a failure. The methodology defines three different workloads ranging from 100GB to 1TB. For details again see appendix 6.4.4.

Parallel Streams

Only tools supporting, or emulating, data transfer parallelism have to be considered. Choosing the number of parallel streams is strictly related with the capacity of memory on both endpoints. A wrong number, e.g. an extremely high one chosen with the idea that more parallel transfers will improve performance, can produce a negative consequence with a significant decrease of the data transfer rate. Thus the methodology considers three different values for the number of parallel stream: 4, 8, and 16.

Qualitative Factors

Qualitative factors, which are not strictly related to the data transfer rate, are also able to provide important information for rating a data transfer tool, like reliability or community acceptance (for more see appendix 6.4.4). The methodology makes use of a 5-level ranking mechanism, from 1 (“really bad”) to 5 (“really good”), along with a short comment or feedback provided by the tester that motivates the evaluation.

Test cases

As a result of combining all different setups for datasets, workloads and parallel streams, the total number of runs for each test is equal to 18. The same test should be executed at least 3 times and the average considered as the final figure. For the list see appendix 6.4.5.

Template

A template for the testing is provided in Annex 6.4.6 and adopted as a common way to collect and present results of tests.

3.2.2 Data transfer tools

The data transfer tests have been performed using the following four tools:

- *UnicoreFTP*, a pluggable file transfer mechanisms provided by UNICORE, available on many PRACE systems [10];
- *GlobusOnLine*, which comes with positive but not exhaustive feedbacks from a preliminary evaluation carried out for Tier-0 systems within PRACE-1IP [12];
- *BBCP*, a tool that is spreading among scientific communities and able to support X.509 certificates for authentication and a data parallelism without requiring a remote server [17];
- *ARC*, a Grid software developed by the NorduGrid and providing data transfer features on top of GridFTP [18]

3.2.3 Testbed definition and preliminary results

All tests results are tracked in a dedicated page of the PRACE Wiki and can be made available on request. This activity will continue in the technology-task T6.3 of PRACE-3IP. After all tests have been finalized the results will be made available in a separate document.

Test benches involved 5 PRACE partners who started testing the 4 tools on the PRACE private network and Internet, as showed in the following table

Tool	Partners	Network
UnicoreFTP	FZJ (Germany), CINECA (Italy)	Internet
GlobusOnLine	CINECA (Italy), EPCC (UK)	Internet
BBCP	CEA (France), CINES (France), EPCC (UK)	PRACE/Internet
ARC	SNIC/NSC (Sweden)	Internet 1/10Gbps

Table 2: Test benches for evaluating new file transfer technologies

As mentioned, a preliminary test phase has been carried out within the available timeframe. A full test phase with comparisons could require several months and the development of script for automating tasks is recommended.

Here we present what has emerged during this preliminary phase.

UnicoreFTP: Tests done between CINECA and FZJ by using the public Internet. First figures showed sufficient results with “Dataset A” (Many Small Files) where a throughput close to the 30% of the maximum available bandwidth was achieved. That is good for a public network where congestion levels are high. Some difficulties have been reported for the software installation and the setup of the environment, along with some problems related to

reliability for long file transfers. It has not been possible to run more tests on the Internet link as well as it was not possible to test the tool against the PRACE network. Despite the ending of the task, activities still go on and now include BSC (Spain) as third partner.

GlobusOnLine: GlobusOnLine provided valuable results already during a similar test made in PRACE-1IP. This is not surprising since it is based on GridFTP. Concerns are still related to security, and in particular to users' privacy because data transfer information is logged on sites that are external to PRACE. Performance is good and generally between 20% and 40% of the total available bandwidth. Reliability is a strong point since no failures have been registered. It has been tested between CINECA and EPCC.

BBCP: Tests of BBCP have been most extensive. The largest benefit of BBCP is the possibility to install it with user privileges and asking for opening a specific port range in the firewall (which can be those already open for GridFTP). It has been tested between CEA, CINES and EPCC on both networks. Even if more tests are needed, the obtained performances have been quite good and similar to GridFTP on the PRACE Network. Reliability, maintenance and fault tolerance have been rated good, too. A further investigation is required in the transfer of dataset type B (Few Large Files), where performance dramatically decreased in a reproducible behaviour.

ARC: ARC has been tested inside the Swedish network of SNIC, which is publicly accessible, by sending data from a local site (NSC) to the SweStore [19], a long-term storage system, on a mixed network made by 1Gbps (for the last mile) and 10Gbps links (carriers). Parallel streams are not supported and only emulated for this test. Another limitation is related to the software footprint, which has been conceived for large Grid environments, which do not fit the typical scenario of PRACE (few large computing systems instead of many small ones). Performance peak has approached 700MB/s, which is 70% of the total available bandwidth of the academic national network.

3.2.4 *Outcomes and lessons learnt*

The following key findings are the results of the preliminary tests executed during this period:

Testing data transfer tools require time. Testing data transfer tools imply considering several factors that are not strictly related to the software solution being tested. Such factors are related to a proper configuration for both sender and receiver hosts, as well as considering networking and security implications. Many of these factors have affected the execution of tests with a resulting delay for producing results and, in some cases, inconsistent figures due to non-optimised configurations.

Tests are also diagnostic activities. As mentioned above, independently from a specific tool, testing bulk data transfer between PRACE sites is a valuable and efficient way to discover possible bottlenecks on the network path and in general incorrect configurations on all "actors" participating in a file transfer, including routers and firewalls. So a good throughput is not the only valuable result, also bad results can provide valuable feedback when indicating some misconfiguration.

GridFTP is still the leader. Even though only preliminary tests have been executed, what has emerged is that it was truly difficult to get better performance than GridFTP. The term performance does not include only data rate but also reliability. This means that work started in DEISA and followed in PRACE [7] [8] for improving user interaction with GridFTP, and spreading its use, are indeed really appropriate. Tests of GlobusOnLine confirmed also excellent responses on this direction.

Dedicated hosts for Data Transfers. In order to provide a reliable, sustainable and high performance data transfer service, it is suggested to consider dedicating specific systems for inbound and outbound data movements. This is because specific host configurations can affect other production services and also because the amount of data to be transferred is even more huge and it really needs dedicated hardware for load balancing. This feedback mainly addresses a common PRACE strategy on big data (c.f. chapter 3.1).

More tests are needed. It is strongly suggested to make a follow-up activity in PRACE-3IP by reusing the same methodology presented here and all efforts spent in setting up the different test benches. It is recommended moreover to extend tests by allowing all involved partners to test all tools against GridFTP.

3.3 iRODS – integrated Rule Oriented Data System

This sub-task follows the iRODS evaluation initiated in the DEISA project. Its scope was to evaluate the recent tool enhancements and to assess the current user needs. To achieve its goals, the work was split into two separate sub-sub-tasks which are described in the following paragraphs.

The goal of the first one was to work on detailed technical evaluations of the current iRODS-release (3.2) which was made available in September 2012. In this context the iRODS User Group Meeting 2013 from February 28th to March 1st in Garching was attended. The agenda and presentations can be found here [36].

The second one focused on the information dissemination (workshop), on a large contribution to the data strategy working group and on pushing external collaborations such as with the EUDAT project. In that workshop the contacts to the developers have been strengthened and the planning for the development could be influenced. Furthermore, users have been instructed on the possible use of these tools for their future data management.

Another concern of the iRODS-sub-task was to provide a methodology during the technical evaluation phase. For this reason a “Feature and Software Evaluation Template” has been defined which was used to provide a homogenous way to lead the evaluation process. This template is used for evaluations found in the appendix 6.6 *iRODS Evaluation Forms*.

3.3.1 Technical evaluations

The technical evaluations have been performed on a testbed set up between five sites, while the details of the testbed can be seen in the Table 3:

Site	1 Gbe Internet Address	10 Gbe PRACE Address	Port	Zone	iRODS Version	Resources
IDRIS	irodsidr1.idris.fr		1247	IDRIS	3.2	demoResc(default)
CINECA	irods-dev.cineca.it		1248	CINECA	3.2	cinecaData(default)
CINES	service4.cines.fr	jade-prace.cines.fr	1247	CINES	3.2	cinesData(default)
NIIF	irods01.niif.hu		1248	NIIF	3.2	niifData(default)
IPB	irods.ipb.ac.rs		1247	IPB	3.2	demoResc(default)

Table 3: iRODS-testbed characteristics

Detailed characteristics of each system and the details of the technical evaluations can be found in the appendix 6.6 *iRODS Evaluations* consisting of seven sub-sections, which report about the specific set of features evaluated (in parenthesis the site responsible for the evaluation is specified):

- 6.6.1 iRODS: Workflow-Objects Evaluation (IDRIS)
- 6.6.2 iRODS: PAM-LDAP-Authentication Evaluation (NIIF)
- 6.6.3 iRODS: Ticket-Based Access Evaluation (CINECA)
- 6.6.4 iRODS: FUSE Evaluation (CINECA)
- 6.6.5 iRODS: Performance Evaluation (CINES)
- 6.6.6 iRODS: Direct-Access Resources Evaluation (IPB)
- 6.6.7 iRODS: iDrop Evaluation (IPB)

The following paragraphs summarize each of these evaluations.

Workflow-Objects evaluation summary

The workflow objects feature provides mechanisms helping users running iRODS workflows in an integrated environment. Although the use of a parameter file allows running workflows in a different context and provides a way for the users to in principle easily interact with the iRODS environment, this feature is currently difficult to use and to integrate in the users' development environment. Thus, users will likely hesitate to use it in the current form.

PAM/LDAP-Authentication evaluation summary

Using PAM, iRODS can be configured to interact with various authentication systems instead of the usual iRODS password authentication. The PAM/LDAP Authentication feature connects with existing PRACE LDAP Authentication. The password exchange is protected with SSL and for subsequent iRODS-commands an iRODS-generated short term (two weeks) password is used. This is stored encrypted in the `.irodsA` file on the client side, which must be protected to assure that an impersonation attack is not possible. Usage of this feature eases the work with iRODS in the PRACE environment substantially.

Ticket-based Access evaluation summary

The ticket based authentication proved to be a very useful feature for short term data sharing, easy to use and reliable. No particular security concern is foreseen.

FUSE evaluation summary

The FUSE [38] module works properly, even if it is not so easy to install. It is used in production environments around the world and proved to be quite useful, even if it is not absolutely reliable: it could be sometimes necessary to umount and remount the collections because sometimes the mount process freezed. Since it provides the users with POSIX-filesystem access methods to their data it seems to be a very attractive feature.

Transfer performances summary

Tuning is always an important part of the work when talking about performance. Once the network was correctly configured, the iRODS tuning part was simple and easy to do and iRODS was then able to provide good performance with the default settings.

iRODS appears here as a simple tool for transferring files allowing good performance. Performance tests must be continued including additional test cases on the PRACE high performance 10 Gbit/s dedicated network to show the full iRODS performance capabilities, since the testbed characteristics mostly did not benefit from the dedicated PRACE network. But the performance figures already seen prove iRODS as being an alternative to the standard transfer tool GridFTP, even offering additional powerful functionalities.

Direct-Access Resources evaluation summary

The Direct-Access Resource feature provides a way to have direct access to the files in an iRODS-resource through the filesystem they reside on. However, taking advantage of this feature depends on iRODS-users having identical userids on the machines that host the

filesystem, and having sufficient file access rights. If the access rights are lacking for a given user, the files on the system will be owned by the root user, and thus not being accessible. Furthermore, since the iRODS server must run as root for the direct access resources to work, the need for this feature should be carefully weighted against possible security concerns.

iDrop Graphical Client Interface evaluation summary

The focus has been on iDrop features and user experience and not on the setting of the testing environment or its integration with iRODS and possible technical issues.

The *iDrop Desktop GUI* is a useful tool, but it still has a lot of place for improvement. The main problem is lag, most likely due to network latency, which doesn't happen when browsing local files. Because of those delays, usage can be quite difficult. There are also only limited search and authentication options. The *iDrop Web Interface* is fast, intuitive and easy to use. It also has limited search options and some authentication problems, but it can include direct links to the *iDrop Desktop GUI* to extend its functionality. To get the most out of the iDrop functionality, Desktop and Web interface should be used together. The lack of support for GSI limits its use in PRACE. Future versions solving these deficits would make this GUI for iRODS an obvious choice for both end-users and administrators.

3.3.2 iRODS Workshop

A workshop has been organized by GENCI/IDRIS-CNRS and SNIC/LIU from September 26th to 28th 2012 in Sweden. The goal of this workshop was to gather people from a wide range of disciplines interested in data management, to discuss users' needs and requirements, to train users and computing centers staff and to tackle the iRODS strategy and future. The following speakers have been specially invited: Prof. Reagan Moore (DICE-UNC), and Leesa Brieger (RENCI-UNC) as well as Jean-Yves Nief from CNRS/CC-IN2P3.

It has been a very successful workshop with more than 30 participants including developer team members, computing centers staff and end-users. The iRODS workshop agenda can be found at [35].

The major data management needs and requirements that have been discussed during the workshop are:

- sharing
- preservation
- data transfer
- replication
- metadata management
- publication
- data mining/ workflow
- storage
- data volume
- EUDAT collaboration

3.3.3 Involvement in the Data Strategy working group

Several partners involved in the iRODS task have been also involved in the *Data Strategy* working group in order to generate a data survey questionnaire aimed to better understand the user needs and requirements regarding the data management within the PRACE project. Some of these partners have been working for PRACE only; most of them have been working both

for PRACE and EUDAT projects. Several topics in this questionnaire such as the data sharing may highlight the iRODS benefits for users compared to a simple data transfert tool.

User answers to this questionnaire as well as finding iRODS pilot projects will define if it is worth deploying iRODS within PRACE as general service. This questionnaire (c.f. section 3.1) can be found in the annex 6.3. When the questionnaire will be submitted to the end-users is not yet decided.

3.3.4 *EUDAT collaboration and pilot projects*

The data management workshop organized by SURFsara from November 26th to 27th 2012 in Amsterdam was attended. As outcome, a pilot project based on the iRODS technology was defined in the scope of the PRACE/EUDAT/EGI collaboration. Since this pilot switched to another technical solution, due to a technical issue identified by EUDAT people, no further work was performed for this pilot by this task. But as response to the cause of that decision, a technical solution was developed and documented. This document describes how PRACE and EUDAT infrastructure can communicate using iRODS by solving the double network interface issue. It is available on the internal documentation server and can be made available on request.

3.3.5 *Conclusions*

From the experience gained in this task, the following future work can be identified:

- get responses from users to the Big Data Survey questionnaire to clarify the applicability of iRODS for PRACE
- combine the questionnaire responses with a future PRACE data strategy
- consider the existing collaborative data infrastructures to develop potential collaborations
- deploy a concrete PRACE/EUDAT collaboration and define the interfaces between both projects
- identify in which context and in which way, iRODS could be deployed in PRACE
- try to influence the iRODS developments as needed
- In the case where relevant use cases are identified:
 - work close enough to the users to understand their data requirements for their entire project
 - offer to the users a long term, community and project oriented solution to their data management question
 - start with a reduced iRODS infrastructure and then build up a full operational infrastructure providing high availability as well as a well a defined iRODS environment for each scientific project
 - define the entire system architecture
 - specify the data management policy
 - specify the conventions to be used
 - define the core services (users and system oriented), access modes and interfaces to the infrastructure

Since iRODS is a technology user communities will utilize for their data management, as seen in projects like EUDAT, it is most likely necessary to support it in PRACE, too. Also the future development should be influenced, to address special HPC-requirements. Thus this activity is planned to be continued in T6.3 of PRACE-3IP.

3.4 File System Technologies

The sub-task *File System Technologies* is part of the task T10.2 *Evaluating data services*. In this section the basic framework for the evaluation of distributed file systems is described. From possible use-cases the technical requirements are derived. Then a methodology for testing, the measurement metrics, initial results and some conclusions are presented.

Four file systems (Coda, Gfarm, Ceph and GlusterFS) have been evaluated. These differ greatly in features, maturity and operational difficulty, but most of them seem to fulfill the requirements for being used as file system being shared between HPC systems.

3.4.1 *The Use-case and the Purpose of the Evaluation*

A common use-case would be sharing user specific personal and configuration data among HPC systems. This would allow users to change HPC sites between DECI calls more easily if they have to, since data would be accessible from more than one HPC site (c.f. section 3.1). Such functionality could also reduce the need for user initiated data transfers, leading to several copies of the same data on different locations.

Another possible use-case could be to give the users a common home directory – same on all sites – with some preconfigured scripts and configuration files to provide them a very similar environment on every PRACE system. This offers the possibility of reviving the initial DEISA philosophy, where users could maintain just one home directory shared on the HPC systems.

3.4.2 *Technical Requirements*

The previously described use-cases require the file systems to offer the following features:

- **Distributed and replicated:** This is needed for moving the users' data from the site where it is created to another automatically.
- **Clustered:** It needs to be mounted on many machines and used by a large number of users concurrently, so it must have cluster functionality built-in.
- **Fault-tolerant:** It has to handle partial downtimes; when only some of the partner sites are unavailable, the users should still be able to use their data on the other HPC systems.
- **Parallel:** It must be scalable by being able to add nodes and storage to the cluster.

3.4.3 *Search Phase*

Scanning through available documentation it has been searched for file system software providing the above mentioned technical requirements. Based on these findings the following file systems have been selected for further evaluation:

- Ceph
- QFS
- Gfarm
- GlusterFS
- Lustre
- Coda
- XtreamFS

QFS [46] and Lustre [45] documented that their architecture has a single metadata server, which cannot be clustered or replicated, so they do not meet the fault tolerance requirement

and have therefore been excluded. The others appeared to be suitable according to their documentation ([40], [41], [42], [43], and [44]) and have been selected for further evaluation.

3.4.4 *Test Environment*

Since it is difficult, if not impossible, to create a test environment on a HPC production system, a non-HPC test environment with a close-by storage system has been established at NIIF. This should have reduced most environment-related effects, like lags based on network latency.

NIIF operates a Fujitsu Systems DX90 S2 Storage box directly connected to the facilitated computing hosts. These hosts contain Intel i7 processors and have 12G memory. KVM is used for the virtualization of the respective cloud infrastructure. The virtual servers, running the latest Debian Squeeze, have been created with eight dedicated storage volumes. Each of these volumes consists of 100 GB on SATA disks configured as RAID6.

The VMs have been configured into two storage nodes and one client.

3.4.5 *Deployment and Feature Validation Testing Methodology*

Identical deployment and feature tests have been applied for each file system, to achieve as much comparability as possible. The following common steps apply to all scenarios:

- Detailed inspection of the installation documentation
- Installation of the latest Debian package available for the given file system (from Debian repositories or from the developers themselves)
- Setup of the software for a two node cluster and a single node client according to the instructions in the manual
- Mounting of the file system on the client
- Troubleshooting any possible problems and making notes of any difficulties
- Proceeding with the evaluation, if everything worked so far and the client can read from and write to the file system
 - Testing replication
 - Writing data once and checking for readability on both nodes
 - Writing data when only one node is online, and observing the repair process when the other node comes back online
 - Testing fault tolerance
 - Testing what is happening if one node from the cluster is becoming unavailable while writing data to it

The results for each step have been documented, which is summarized in the next section.

3.4.6 *Test Results*

Finally four file systems could be tested according to the aforementioned methodology.

Ceph

Ceph is well documented and the software is easy to install, because the developers themselves make Debian packages from the latest versions of the software. A quite complex initial configuration was required, but then mounting on the client was an easy task.

Replication worked out of the box. Technologically the replication is based on a consistent hashing algorithm and synchronized replication, so the client itself is informed by the server

how many replicas are needed and where to place them; it does not get a 'write successful' notice back until all the needed replicas are written to the nodes. This results in a tradeoff for the write performance but assures a high availability in a multi-site setup.

The automatic repair function worked, too. The nodes synchronized data after the node with the outdated data joined back. Fault tolerance was seamless; the client did not even notice a server node going offline when it was writing data to it.

Although fulfilling every aspect of the feature validation tests, the developers themselves state that parts of the file system are not yet stable enough to be used in production environments. Furthermore the complexity of the configuration has to be noted.

GlusterFS

GlusterFS is also well documented and has stable Debian packages available from the official repository. The file system configuration has been straightforward and it could be mounted on the client easily.

The replication worked seamlessly when both nodes have been online, but the write failover from one node to another, when a server node went offline during a write, stalled the client for more than 10 seconds. The write operation was in a deadlock state in the operating system until a timeout occurred, but then it continued successfully. The automatic repair was also working successfully after the offline node joined back online.

Thus, GlusterFS is pretty simple to install and operate, all the required features are available, but it lacks some agility and tuning options.

Gfarm

Gfarm has an outdated and incomplete documentation and the Debian packages provided in the repositories are built from non current versions. Based on such old software a two node file server environment has been successfully setup, but the client could not connect to them and thus not mount the filesystem. This has been identified as bug [47] which is already fixed in a newer version.

Therefore, it was attempted to compile a current version from the source. There have been several header and library errors which could not be fixed.

So it was decided to postpone Gfarm evaluation and no result can be presented yet.

Coda

Coda has an extensive documentation but some of it is referring to older versions. Since the developers supply Debian packages for Coda software installation went smoothly. The configuration was difficult due to the inappropriate documentation. Finally the two node cluster exported the file system to the client where it could be mounted successfully.

Although stated by the documentation as available, the replication-feature did not work. Further investigation is required to determine if this is only due to a wrong configuration.

3.4.7 Detailed Description of the Performance Measurement

Two major questions related to file system performance are of special interest. First, what is overhead introduced by a distributed file system compared to a local one? Second, how does the distributed file system scale and how does this influence the performance? This finally will allow for the comparison of the different distributed file systems.

Specific FIO [48] jobs have been used to get answers to these questions, where each of the file systems had to undergo all the tests described in the matrix below:

I/O type	operation type	block size		
		4KB	16KB	32KB
synchronized	read	X	X	X
	write	X	X	X
	read + write	X	X	X
asynchronous	read	X	X	X
	write	X	X	X
	read + write	X	X	X

Table 4: Matrix of test cases for file systems

A simple 4KB synchronized read FIO job for the raw block device looked like this:

```
[random-read-sync-4K]
rw=randread
size=512m
blocksize=4k
directory=/mnt/ext4
```

In addition, four more complicated workload simulation jobs with multiple read and write operations at the same time running in a parallel way, have been executed. They consisted of several such job-definitions running in at the same time utilizing different parameters each.

All the tests have been first run on the raw block device provided by the storage box, then on a local ext4 file system on top of this block device. These numbers served as baseline for the comparison with the distributed file systems. Finally, the entire test suite was run for each of the distributed file system selected. Every run was executed 10 times. Taking the average of the results of these 10 runs should have eliminated any possible jitter.

FIO returns about 60 different measurement values when finishing. The results can be categorized as follows:

- For read and write operations individually
 - Completion latency
 - Submission latency
 - Bandwidth
- For the job globally
 - CPU usage
 - I/O depth distribution
 - I/O latency distribution

All of this data was collected into spreadsheets out of which more important and interesting summary-information was extracted. These findings will be shown in the next sub-section.

3.4.8 Performance Measurement Results

In this sub-section a selection of results is presented documenting some general conclusions which could be deduced from the measurements already. Furthermore, two figures visualize the general findings.

It was found out, that even on a local file system (here: ext4) the random read performance can be lowered by a factor of 3 or in extreme cases even 10 compared to a raw block device

for synchronized and also asynchronous operations alike. The reason for this is most likely a lot of inode-related operations before accessing any data-part of a selected file. The random write performance is mostly not affected and its speed on an ext4 file system is about 60-80% of the speed of the raw block device for synchronized as well as for asynchronous operations.

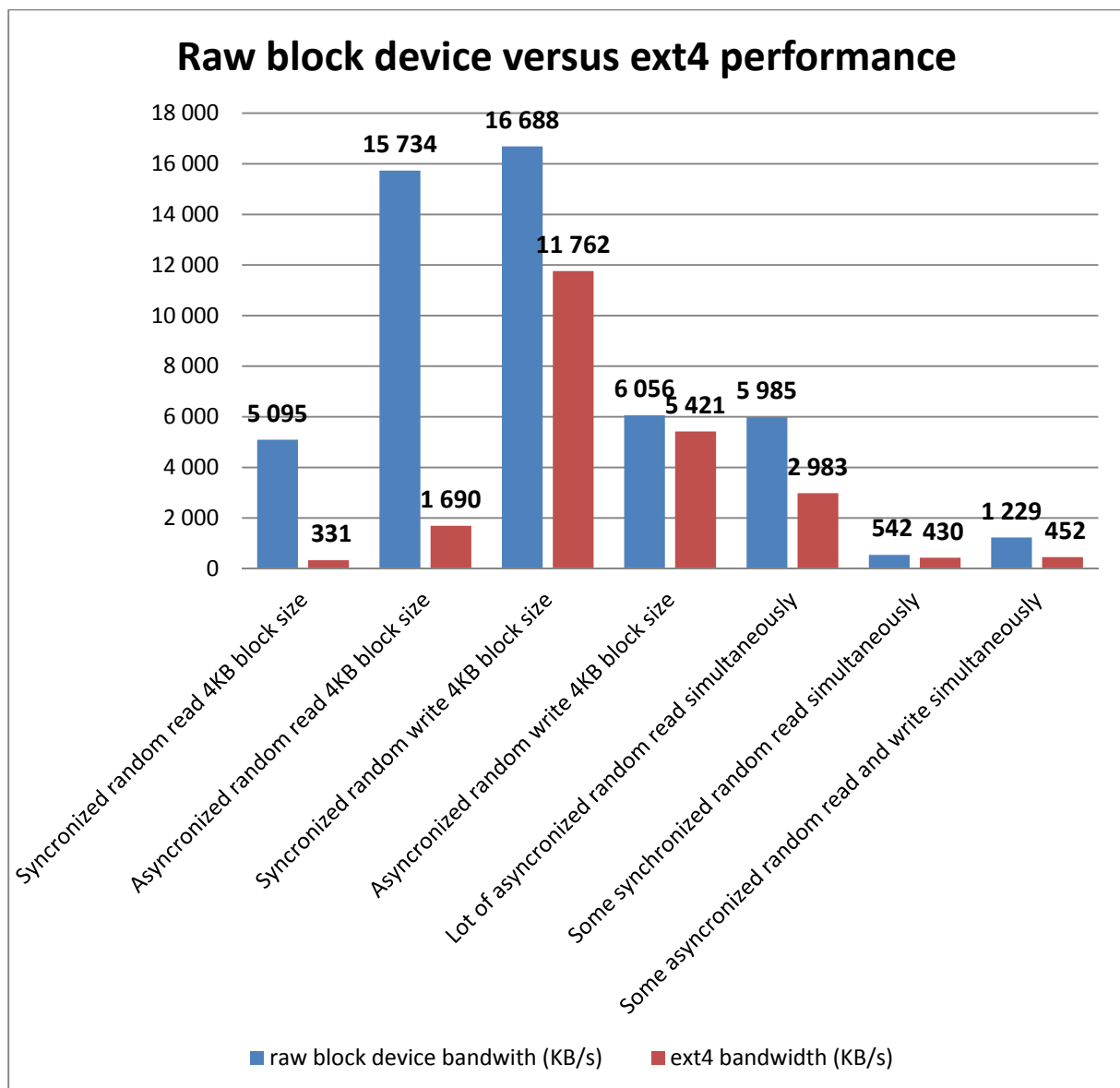


Figure 2: File System Performance Comparison (Raw Blocks)

In most cases introducing a distributed file system to the setup does not further decrease the random read performance, but in fact in most cases even raises it by a factor of about 2, since the data can be retrieved from two locations/servers. This is a perfect example of the read performance benefitting from a distributed file system.

This advantage turns to the opposite when looking at the random write performance of a distributed file system. It is slowed down by a factor of 5 to 9 compared to the ext4 file system. This drawback most likely can be relieved by adding much more nodes to the cluster than the target number of replicas, because in that case a distributed file system can stripe writes well across multiple nodes for a better overall performance. This assumption should be investigated further in a later phase of the evaluation.

The other interesting finding was that all distributed file systems show comparable performance within a range of less than 40% deviation.

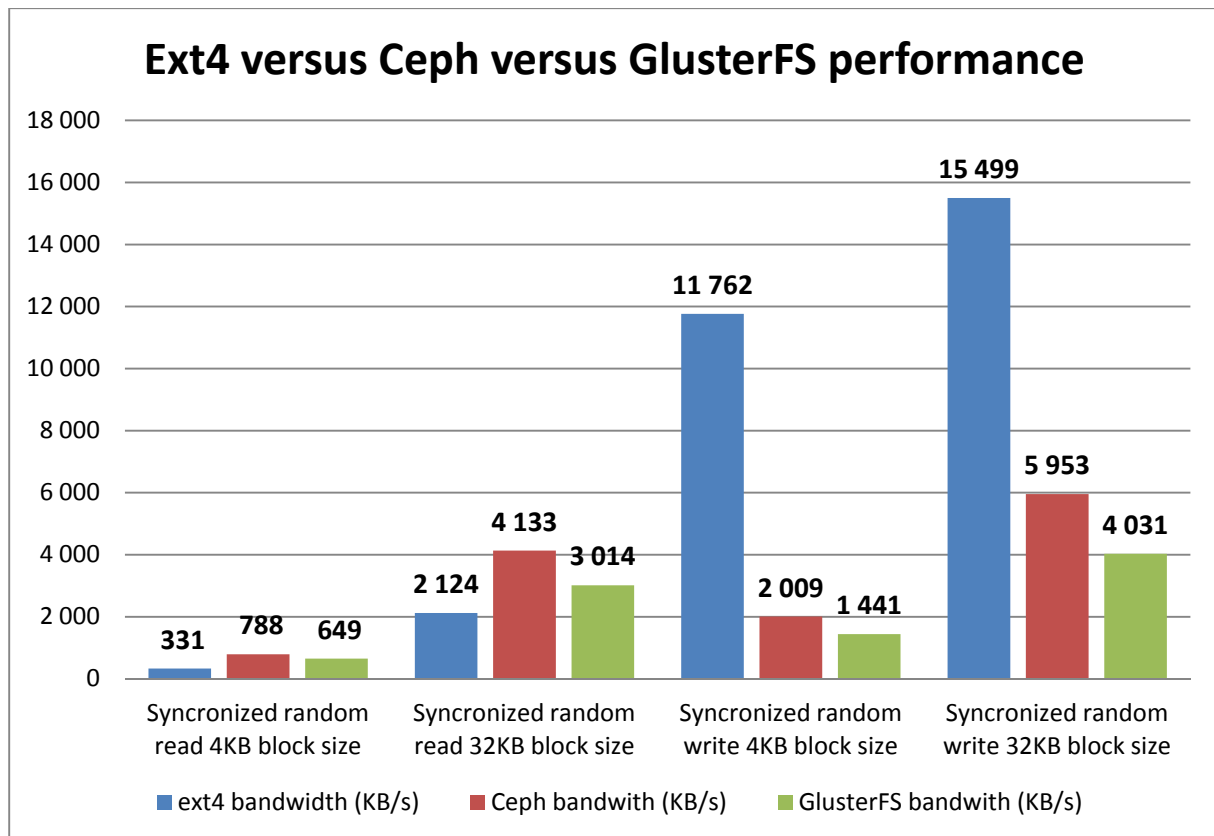


Figure 3: File System Performance Comparison (Ext4, Ceph, GlusterFS)

3.4.9 Conclusions and Plan for Further Work

The use-cases for possible common home-directories or easily accessible shared data spaces across HPC systems in PRACE illustrate the general usefulness of distributed file systems. The tests of the different software solutions concerning reliability and performance do not yet indicate clear recommendations. But since the future data strategy in PRACE is still open, it seems to be wise, to be prepared for eventual requests for the provision of a shared storage based on a distributed file system. Thus, it is planned to continue the work in the task T6.3 of PRACE-3IP, which is scouting technological developments of potential interest for PRACE.

Thus, Gfarm and Coda should be made working properly for testing their behaviour. XtremFS, and further ones, like OrangeFS and FraunhoferFS, missed in the search process, should be included in an authoritative comparison. Furthermore, the scalability and multi-site usage should be tested with more partners on the PRACE distributed infrastructure.

4 Remote Visualization

4.1 Introduction

As stated in deliverable D10.1 [3], the work regarding remote visualization solutions, systems and services has mainly focused on the class of solution that are application transparent (as much as possible) and session oriented (so each users own their visualization sessions). Those solutions are mainly represented by VNC-like systems.

Among the different available VNC solutions reported in the previous deliverable, PRACE centres have relied on TurboVNC/VirtualGL open source solution for deploying visualization

services over WAN, offering remote visualization services even at researchers at home, connected with consumer grade ADSL lines.

Each partner has organized its visualization service using different hardware and adopting different access policies (queued sessions, advanced reservations, special (reserved) visualization nodes) but all used the same underlying technological platform using the VirtualGL project for application neutral OpenGL remotization scheme and TurboVNC as the VNC-server/client component.

SURFsara has investigated a number of remote visualization topics that are of interest due to current trends in computing and visualization. A first topic was the use of VirtualGL/TurboVNC for high-end high-resolution large screen visualization setups. Secondly, an investigation was made into the possibilities of using GPU-compute hardware for remote visualisation, together with a comparison between GPU-based rendering and software rendering. The detailed results will be published in a separate PRACE whitepaper.

CINECA had used a proprietary VNC technology from IBM (DCV) to support technical users that need specific proprietary visualization applications in engineering and flow simulation (StarCCM, Fluent, etc.). The DCV technology is currently provided and supported by NICE and is still in use as an embedded component of a customized web portal for access to technical computing resources based on NICE EngineFrame

SNIC/LiU explored other remote visualization technologies available and investigated deeply into the Teradici PcoIP solution described below in section 4.2. It can be used when top performance or complete application transparency were needed and a high speed, low latency, campus wide network backbone was available.

The second year focused on the evaluation of the performance of the different VNC based services under different usage conditions (see section 4.4) and the further development of the CINECA RCM [52] pilot project, which aims at the simplification and the improved deployment of the TurboVNC/VirtualGL [51]/[50] software stack and is described in more detail in section 4.3.

In this context RZG has tested the CINECA RCM pilot from the applications and operations point of view. RZG staff has compared the user's experience of CINECA RCM with that of a standard VirtualGL/TurboVNC-based solution which is operated by RZG for the MPG and which has been made available also to PRACE users for analysing their simulation data produced on RZG-system in the context of DECI projects. RCM was successfully tested (using a Paraview application example) with client software for the operating systems Ubuntu 10, OpenSUSE 11, and Windows 7. According to RZG's experience, RCM addresses some of the shortcomings of standard VirtualGL/TurboVNC-based solutions. In particular, RCM provides a more convenient way to reserve and access remote visualization resources. Specifically, reservation of resources and tunneling through firewalls is handled more transparently and in a more user-friendly way by RCM. RCM also allows sharing GPU and CPU resources and thus enables “virtualizing” the resources.

4.2 Teradici PCoIP setup at SNIC/LU

Teradici PCoIP technology [53] enables efficient and secure transfer of pixels including associated session information (such as mouse, keyboard, USB and audio) across a standard IP network. It provides full frame rate 3D graphics and high-resolution media.

The PCoIP protocol encrypts and compresses the data stream on the server side using either dedicated hardware or in software (using VMware). The data stream is received and decoded

at the receiving end using a stateless¹ "zero client" or in software (VMware View). The software solution does not currently support Linux as host operating system. The latest generation stateless device supports up to two channels at 2560x1600 or four channels at 1920x1200 and includes VGA, DVI and DisplayPort display interfaces.

The hardware-based solution is 100% operating system and application independent. The video signal from the graphics card is routed directly to the PCoIP host adapter where it's processed using hardware and transferred to the network using the onboard dedicated GigE NIC. Power, USB and audio are handled over the PCIe bus.

The SNIC hardware based PCoIP solution consists of two dedicated graphic nodes that is part of the production HPC cluster "Alarik". The graphic nodes have 32 GB RAM, 16 cores (2 sockets) and Nvidia Quadro 5000 graphic cards. Each node is equipped with an EVGA PCoIP host adapter card that ingests the pixel stream(s) from one or both DVI-D outputs of the Quadro 5000 card. On the client side currently two different appliances are used; an EVGA PDO2 "zero client" and a Samsung 24" monitor with integrated PCoIP client i.e. the monitor connects directly to the Ethernet socket.

The current setup is point-to-point and serves "power users" at the campus with a high performance, secure remote visualization mechanism. No long-distance WAN tests have been possible to perform.

Main application area is post processing of large CAE data-sets using software such as Abacus CAE and Paraview. From a user experience it is equal to using a local workstation with respect to authentication and usage but of course much more powerful since the system is an integrated part of the computational cluster. Then main operating system in use is Centos but one of the visualization nodes has been running MS Windows as part of the test.

An important benefit that distinguishes this setup from software-based solutions is the remote visualization subsystems independence from the host computer as described above in further detail. No specific software or drivers need to be loaded and hence there is nothing that might conflict with the operating system or end user applications.

Furthermore the solution puts no additional load on the host such as CPU cycles needed for image compression, host to graphics bandwidth for image readback, etc. This allows the application to run at full speed as if displayed to a local monitor. Achieved remote image quality is only determined by available network performance.

The possibility to enable secure USB bridging to the host system opens up interesting options for transferring data and connecting other (interaction) devices. An administrator can disable this function if needed.

PCoIP is a commercial solution using proprietary hardware both on server and client side, something that somewhat limits the usage for academic purposes even if the price level is very decent, especially when put into a performance and image quality context.

Performance-wise the resulting image quality and interactive performance is perceived as very good and predictable when running on the campus network using a 1920x1200 resolution. The technology adapts to different network situations in a user-controllable fashion to allow either automatic adjustments or using fixed numbers such as maximum peak bandwidth allowed and how the system should behave during congestion.

¹ Stateless means there is no record of previous interactions and each interaction request has to be handled based entirely on information that comes with it. PCoIP encodes each pixel to a lossless state once they stop changing to ensure a pixel-perfect image.

The bandwidth needs depend on the frame content, spatial resolution, number of display channels and other communication, such as audio and USB. The largest contribution to the bandwidth usage is the portal pixel transfer, others (less contributing) are audio, USB bridging and to an almost negligible extent, system management. Network latency up to 150ms are supported and responsiveness typically gets sluggish around 40-60ms. This is however subjective and session dependent.

4.3 CINECA Remote Connection Manger

The Remote Connection Manager CINECA pilot project has already been described in an annex included in the previous deliverable D10.1 [3].

The system is available since almost one year on the CINECA PLX cluster nodes and has been recently enhanced to support new graphics nodes, different access modes and has also been used to support non-accelerated VNC sessions on front end nodes of CINECA Blue Gene/Q Tier-0 machine.

The client part consists of a single executable that wraps the TurboVNC client and the python code dealing with ssh tunneling, needed to support visualization services installed in compute nodes that are not directly accessible. The client supports re-connection to open sessions and PAM authentication. It does not handle session sharing or VNC password. The client is able to auto-update when a new version is available.

The server-side currently supports session book-keeping and has support for PBS (PLX cluster), LoadLeveler (Fermi BG/Q), as well as direct ssh-access. The code is available from the web under <https://hpc-forge.cineca.it/svn/RemoteGraph/trunk/>.

The service has been tested with different open-source visualization applications such as ParaView [54], Blender [55], Visit [57], OpenCV [58], MeshLab [56], and others. It supports pre-compiled codes as the UniGine [59] graphics engine test as well as pre-compiled ParaView deployment, but there have been some issues with StarCCM [60] visualization code.

4.4 Performance evaluation of VNC based remote visualization services

In all visualization applications the overall satisfaction of the user interacting with the system is the most relevant criteria for the evaluation of the system. Therefore the most important parameters for the evaluation are those effectively perceived by the user:

- the effective frame-rate at client side
- the overall latency of the system
- the visual quality of the image stream

It is important to underline that these parameter must be measured taking into account all the components that compose the client-server system:

- Server side hardware platform (CPU / GPU)
- Application code
- OpenGL interposition layer (VirtualGL)
- VNC image compression (TurboVNC server)
- Network transport (depends heavily on network bandwidth)
- VNC client platform for image decompression and stream rendering

It was decided to concentrate on the frame rate parameter as the other two, even if very important in determining the overall user satisfaction, are much harder to estimate in a quantitative way.

Almost all the VNC clients use aggressive lossy image compression schemes to trade off image quality for frame rate, usually on single images as the more effective interframe compression schemes used in video streaming generate excessive latency. However, this loss in image quality is really difficult to measure in a quantitative way as it heavily depends on image content itself.

In order to quantify latency, a proper setup is needed (high speed camera) and the procedure can be significantly time consuming, as described in an article *The truth about latency in Cloud Gaming* [49]; furthermore, since latency is mostly dominated by the network components, it can be highly variable depending on the client-server network load.

In order to quantify the frame rate, a tool (*tcbench*) included within the VirtualGL distribution, which adopts a simple but effective approach, has been used. The tool runs on the client machine and inspects a small portion of the VNC window detecting how many times the screen changes per second. If an application is run which constantly changes the screen, then the tool correctly detects the screen change and computes the real perceived frame rate, disregarding frame spoiling techniques.

Regarding which application is used for testing, two approaches are possible: the first is to use a very simple (and fast) graphic application to minimize the application overhead to be sure of being limited by just the grab-compression-transport-decompression involved in remote visualization.

Another approach is to use a graphic application that is able to render enough frames to saturate the image transport layer but is nevertheless representative of a real application with sufficient image complexity and variance.

For that purpose a demo of a graphics engine that pushes the limits of our old GPU but run smooth on new ones has been executed. The tests have confirmed that the default settings that TurboVNC defines for the image compression setup are indeed the most appropriate for LAN as well as for high speed WAN as with them TurboVNC exhibits very few compression artifacts (almost unnoticeable) and optimizes all other costs as well as frame rate.

Depending on available bandwidth, it could be necessary to adopt more aggressive image compression settings in order to make use of the full GPU power available to attain a perceptual satisfactory experience.

The next figure shows from left to right the same image as a sequence using lossless zlib, lossless jpeg, and default settings; there is almost no noticeable artifact.

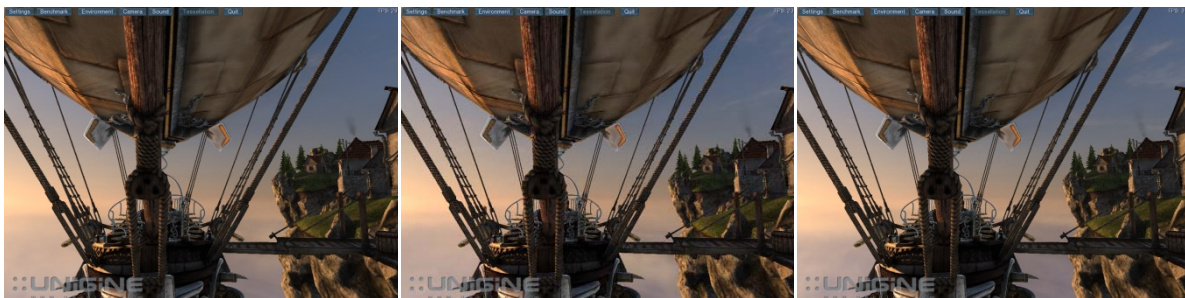


Figure 4: Images compressed with lossless zlib, lossless jpeg, and default settings

While the next figure again from left to right shows the sequence with jpeg compression suggested for WAN, custom compression set to 12%, and custom compression set to 7 %.

The two latter compression factors cause really annoying artifacts. Thus testing was limited to the 12 %, since asking for more compression resulted in unbearable artifacts.

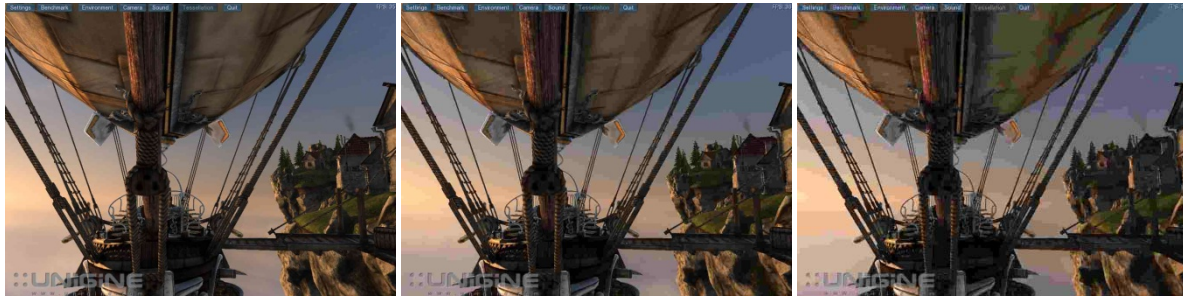


Figure 5: Images with jpeg compression with WAN setting, 12%, and 7% setting

The RVN UniGine tests show that there is no gain in optimizing image compression when the frame rate bottleneck resides on the remote GPU resources; they also show how the same application can hit different limits when different resources become available: applications that require most server side resources are the ones that most benefit from a remote visualization service.

It must also be noted that there is a non negligible load on the login node for the ssh tunnel execution in the visual queue UniGine tests: this load seems to be connected to the raw volume of data transfer, so directly related to the available bandwidth used, which is in turn related with the image compression schema adopted and the frame rate attained. Nevertheless, in VNC sessions performing image transfer at full speed, the load on the login node can be up to one-third of that imposed on the compute node; this can become an issue in case many visualization nodes are served by the same login node.

More details on the performance tests can be found in the appendix 6.5.

5 Summary and Future Work

All three tasks made good progress towards improving the infrastructure. Several direct benefits for the users, e.g. the web and portal related tasks as well as remote visualization offerings, could be achieved by the first and the third task. Furthermore, the first task advanced the PRACE-internal management of the infrastructure related to accounting, service certification, and monitoring. Valuable input came from the collaborations with other technologically oriented project; here especially the input from user communities in the pilot projects helped better understanding user needs. These influenced particularly the second task. This task is not fully HPC-centric and more long-term oriented, since data-management is not only of high importance for HPC-users. HPC-generated data is also very often further processed outside HPC-systems. Thus, many of the results achieved here do not yet bring direct improvements for the users or the infrastructure, but are of importance for the further strategic decisions of PRACE concerning the handling of data in the future.

As already indicated in the respective sections many of the activities have potential or even needs for further investigation or development. Therefore, a two-day hand-over meeting for the important activities of WP10 to be continued in Task 6.3 of WP6 in PRACE-3IP has been held in Garching near Munich in June 2013. The relevant tasks have been identified and the planning for their continuation has been defined, so the work can continue seamlessly.

6 Annex

6.1 PRACE Event Integration – Screenshots

The screenshot displays the PRACE website's 'PRACE Training events' page for June 2013. The page includes a navigation bar at the top with links like 'Home page', 'Forum', 'FAQ', 'Job vacancies', 'Press Releases', 'Newsletters', and 'Contact'. A search bar is also present. The main header features the PRACE logo and the text 'PARTNERSHIP FOR ADVANCED COMPUTING IN EUROPE'.

The left sidebar contains a menu with categories such as 'About PRACE', 'HPC access', 'PRACE FP7 Projects and Outcomes', 'Events', and 'Media'. The 'PRACE Training events' link is highlighted.

The main content area is titled 'PRACE Training events' and contains introductory text about the PRACE RI and its training goals. Below this is a calendar for June 2013, showing various training events scheduled throughout the month. A dropdown menu is open over the 11th of June, listing options like 'PRACE Seasonal Schools', 'PATC Courses', and 'Partners Trainings'.

The right sidebar features a 'High Performance Computing' banner, a 'Call announcements' button, 'Job vacancies', 'PRACE Press releases', and 'PRACE Newsletters'. It also includes a 'WATCH LIVE NOW!' section for 'PRACE SUMMER SCHOOL 2013' and 'PRACE AUTUMN SCHOOL 2013', along with a 'PRACE News' section listing recent updates.

Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
27	28	29	30	31	1	2
		Code coupling using OpenPalm @ CINES				
		Statistical Analysis for Post-Genomic Data: Parallel Computing with R @ EPCC				
3	4	5	6	7	8	9
Introduction to CUDA Programming @ BSC						
	PRACE Seasonal Schools	11	12	13	14	15
		Partners Trainings				
17	18	19	20	21	22	23
PRACE Summer School 2013 - Frameworks for Scientific Computing on Supercomputers						International Summer School on HPC Challenges in Computational Sciences
	Hybrid MPVOpenMP programming @ IDRIS					
24	25	26	27	28	29	30
International Summer School on HPC Challenges in Computational Sciences						
		Cray Advanced Tools workshop @ EPCC				
1	2	3	4	5	6	7
Shared-Memory Programming with OpenMP @ EPCC		Message-Passing Programming with MPI @ EPCC				

Figure 6: Event Integration Screenshot 1 – PRACE Training Events

The screenshot displays the PRACE website interface. At the top, there is a navigation bar with links for Home page, Forum, FAQ, Job vacancies, Press Releases, Newsletters, and Contact. Below this is a large banner with the PRACE logo and the text 'PARTNERSHIP FOR ADVANCED COMPUTING IN EUROPE'. The main content area is divided into sections for 'ON THE WEB' and 'Upcoming PATC courses'. A sidebar on the left contains a navigation menu with categories like 'About PRACE', 'HPC access', 'PRACE FP7 Projects and Outcomes', 'Training and Documentation', 'Events', and 'Media'. The right sidebar includes a 'High Performance Computing' section with a call to action, 'Job vacancies', 'PRACE Press releases', 'PRACE Newsletters', and 'WATCH LIVE NOW!' for two summer schools. The main content area lists several upcoming courses:

- Advanced Fortran Topics & Coarray Fortran @ CSC** (Monday 10 June 2013): This brand new course gives introduction to more modern features of the Fortran 2008 standard. Among subjects are more flexible handling of dynamically allocated objects, operator overloading, object oriented features and language interoperability. Thereafter a highly useful introduction to (...)
- Cray Advanced Tools workshop @ EPCC** (Wednesday 26 June 2013): Cray's supercomputer platforms are an advanced pairing of software and hardware that provide HPC application developers and users the opportunity of excellent scaling and high productivity. This workshop, provided staff from the Cray Centre of Excellence for HECToR, offers instruction and (...)
- Shared-Memory Programming with OpenMP @ EPCC** (Monday 1 July 2013): Shared-Memory Programming with OpenMP Almost all modern computers now have a shared-memory architecture with multiple CPUs connected to the same physical memory, for example multicore laptops or large multi-processor compute servers. This course covers OpenMP, the industry standard for (...)
- Message-Passing Programming with MPI @ EPCC** (Wednesday 3 July 2013): Course Description The world's largest supercomputers are used almost exclusively to run applications which are parallelised using Message Passing. The course covers all the basic knowledge required to write parallel programs using this programming model, and is directly applicable to almost (...)
- Petaflop Supercomputer @ LRZ** (Upcoming): This workshop gives an introduction to the usage of the new Petaflop class Supercomputer at LRZ, SuperMUC. The first three days of this are dedicated to presentations by Intel on their software development stack (compilers, tools and libraries); the remaining day will be comprised of (...)
- Advanced OpenMP @ EPCC** (Wednesday 17 July 2013): OpenMP is the industry standard for shared-memory programming, which enables serial programs to be parallelised using compiler directives. This course is aimed at programmers seeking to deepen their understanding of OpenMP and explore some of its more recent and advanced features. This two-day (...)
- Advanced Fortran Topics @ LRZ** (Monday 16 September 2013): This course is targeted at scientists who wish to extend their knowledge of Fortran beyond what is provided in the Fortran 95 standard. Some other tools relevant for software engineering are also discussed.
- Fortran 95/2003 @ CSC** (Monday 30 September 2013)

Figure 7: Event Integration Screenshot 2 – Upcoming PATC Courses

The screenshot displays the PRACE website's 'PATC Courses' page for June 2013. The main content is a calendar grid where blue bars represent scheduled courses. A mouse cursor is hovering over the 'PATC Courses' link in the left sidebar. The right sidebar contains various news items and announcements, including 'WATCH LIVE NOW!' for PRACE Summer School 2013 and PRACE Autumn School 2013.

Navigation: Home page | Forum | FAQ | Job vacancies | Press Releases | Newsletters | Contact

Header: PRACE PARTNERSHIP FOR ADVANCED COMPUTING IN EUROPE

Left Sidebar: About PRACE, HPC access, PRACE FP7 Projects and Outcomes, Training and Documentation, Events, Media, HPC logo, Primeur logo.

Main Content: Home page > Training and Documentation > PRACE Training events > PATC Courses

PATC Courses Calendar (June 2013):

Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday	
27	28	29	30	31	1	2	
		Code coupling using OpenPalm @ CINES					
		Statistical Analysis for Post-Genomic Data: Parallel Computing with R @ EPCC					
3	4	Introduction to CUDA Programming @ BSC				8	9
10	11	12	13	14	15	16	
12th Summer School on Scientific Visualization @ CINECA							
17	18	19	20	21	22	23	
Hybrid MPI/OpenMP programming @ IDRIS							
24	25	26	27	28	29	30	
Cray Advanced Tools workshop @ EPCC							
1	2	3	4	5	6	7	
Shared-Memory		Message-Passing Programming with MPI @ EPCC					

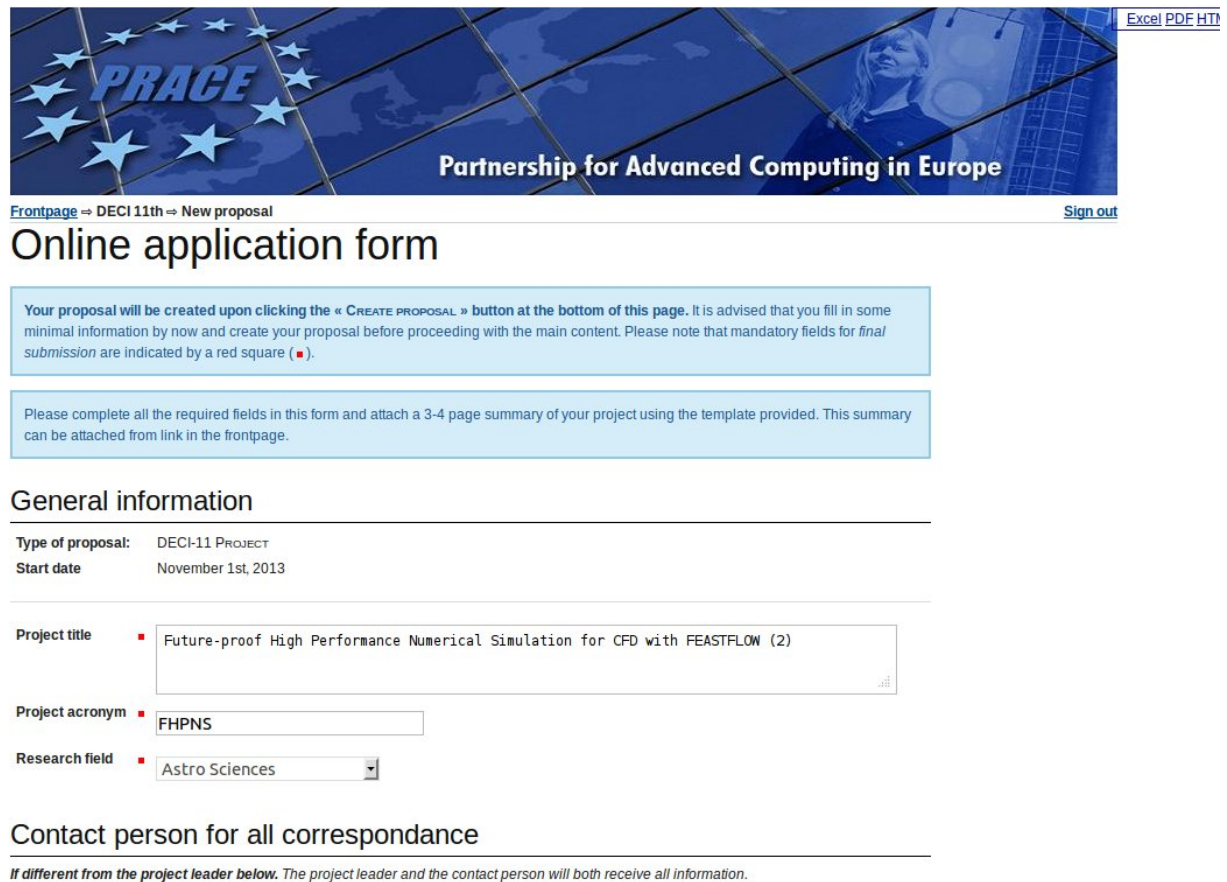
Right Sidebar: High Performance Computing, European Commission, Call announcements, Job vacancies, PRACE Press releases, PRACE Newsletters, WATCH LIVE NOW!, PRACE SUMMER SCHOOL 2013 (17-21 July, Ostrava, Czech Republic), PRACE AUTUMN SCHOOL 2013 (23-27 September, Ljubljana, Slovenia), PRACE News, PRACE Summer School, Ostrava, Czech Republic (18 June 2013), PRACE Educates Tomorrow's Researchers (14 June 2013), Welcome to the PRACE booth at ISC'13 Exhibition (14 June 2013), DECI-11 Call closing deadline extend to 14th June, 17:00! (7 June 2013), The PRACE Scientific Conference with success stories and strategies in European HPC (6 June 2013), More PRACE News, PRACE related news, A coordinated approach for science in Europe (19 June 2013), PRACE Educates Tomorrow's Researchers.

Footer: Home page | Contact | Site Map | Statistics | visits: 400482 | EN | Training and Documentation | PRACE Training events | PATC Courses | OPML | Site created with SPIP 2.1.23 + AHUNTSIC | (C) Copyright PRACE AISBL, 2010, all rights reserved | Hosted by CINES | CC BY SA

Figure 8: Event Integration Screenshot 3 – PATC Courses

6.2 DECI Portal

The next picture shows a screenshot of the PPR-Tool with a setup for the latest DECI-11-call after logging as applicant:



[Excel PDF HTI](#)

[Frontpage](#) ⇒ [DECI 11th](#) ⇒ [New proposal](#) [Sign out](#)

Online application form

Your proposal will be created upon clicking the « CREATE PROPOSAL » button at the bottom of this page. It is advised that you fill in some minimal information by now and create your proposal before proceeding with the main content. Please note that mandatory fields for final submission are indicated by a red square (■).

Please complete all the required fields in this form and attach a 3-4 page summary of your project using the template provided. This summary can be attached from link in the frontpage.

General information

Type of proposal: DECI-11 PROJECT
 Start date: November 1st, 2013

Project title: ■ Future-proof High Performance Numerical Simulation for CFD with FEASTFLOW (2)

Project acronym: ■ FHPNS

Research field: ■ Astro Sciences

Contact person for all correspondence

If different from the project leader below. The project leader and the contact person will both receive all information.

Figure 9: DECI-PPR-Tool Screenshot

6.2.1 DECI peer review tools functionality comparison table

#	Functionality	Rate	HPC-Europa Tool	PRACE Tier-0 PPR Tool	Comment
1	Electronic submission of project proposals.	Essential	Well supported	Well supported	
2	Developers' ability to programmatically redesign the forms contents and their integration with the internal database.	Essential	Well supported	Well supported	
3	Web-based ability (form design tool) to design and change the project submission and evaluation forms.	Desiderata	Well supported	Partially supported	PRACE T0: Planned in portable kernel roadmap.
4	Provide users with complete online control of their data (application form, user data etc.) and enable them to effectively	Essential	Well supported	Well supported	

#	Functionality	Rate	HPC-Europa Tool	PRACE Tier-0 PPR Tool	Comment
	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	Not supported	Well supported	PRACE T0: Roles are implemented. However, Admin UI delegation is not yet available
6	Store applicants' data, project data, TE review data, suggested extra TE info, SE data, ranking info etc. into the DECI Database	Essential	Not supported	Well supported	HPC-Europa: Integration with the DECI database is needed.
7	Create and/or change user's, evaluator's, site's, countries, info.	Essential	Partially supported	Partially supported	PRACE T0: 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	Well supported	Well supported	
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 tracked via the proposed tool.	Desiderata	Partially supported	Well supported	PRACE T0: Should be ok, to be precised. HPC-Europa: general statistics on the entire review process are available though.
10	Copy or link the relevant data from the web-based tool, when	Essential	Not supported	partially supported	PRACE T0: Linking should be possible, with quite

#	Functionality	Rate	HPC-Europa Tool	PRACE Tier-0 PPR Tool	Comment
	needed, into the DPMDB (i.e. project name, home site, technical requirements such as CPU type, number of jobs, memory, simulation codes etc.)				reasonable 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)	Desiderata	Not supported	Partially supported	PRACE T0: 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)	Desiderata	Partially supported	Well supported	PRACE T0: Such features already exist 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.	Desiderata	Well supported	Well supported	
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).	Desiderata	Well supported	Well supported	
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 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,	Desiderata	Not supported	Partially supported	PRACE T0: Included in kernel development roadmap. HPC-Europa: Easy to develop.

#	Functionality	Rate	HPC-Europa Tool	PRACE Tier-0 PPR Tool	Comment
	pending or overdue reviews).				
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	Partially supported	Partially supported	HPC-Europa: Basic tools (e.g. reopening, deadline change, etc.) are already available. Advanced ones should be better clarified. PRACE T0: - Reopening applications for the administrative process: Available. - Changing the deadline of a review: Available. - Changing the discipline category of a project: can be implemented.
17	Communicate to the centers the info of awarded projects (LDAP) "Project ID, User Accounts, etc."	Essential	Partially supported	Well supported	HPC-Europa: LDAP compliant information can be already exported but specific developments could be necessary according to LDAP schema. PRACE T0: Specific export process already implemented for the current schema.
18	Create a report of reviewers, with past historical information (reviews attributed and reviews in previous calls), including passwords	Essential	Partially supported	Well supported	HPC-Europa: Easy to implement. PRACE T0: Already implemented in the administrators access.
19	Create a report of all persons involved in past and present calls (PIs, collaborators) with history (call, proposal ID, ...)	Essential	Well supported	Well supported	PRACE T0: Already implemented.
20	Guarantee a highly secure log-in system (highly secure password)	Essential	Well supported	Partially supported	CINES: Connection in HTTPS, stronger security of password is planned for the next version of the tool.

Table 5: Test benches for evaluating new file transfer technologies

6.2.2 DECI peer review tool functionalities requirements list

#	Functionality	Rate	Implementation status (July 2013)
1	Electronic submission of project proposals.	Essential	Fulfilled
2	Developers ability to programmatically redesign the forms contents and their integration with the internal database.	Essential	Yes
3	Web-based ability (form design tool) to design and change the project submission and evaluation forms.	Desiderata	Mid-2014
4	Provide users with complete online control of their data (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).	Essential	Fulfilled
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	Ad-hoc implementation of user profiles. Missing user interfaces.
6	Store applicants' data, project data, TE review data, suggested extra TE info, SE data, ranking info etc. into the DECI Database.	Essential	Postponed
7	Create and/or change user's, evaluator's, site's, countries, info.	Essential	Postponed
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	Not yet scheduled
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 tracked via the proposed tool.	Desiderata	Not yet scheduled
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	Not yet scheduled
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).	Desiderata	Not yet scheduled
12	Create and export documents and information that should feed other systems or processes (i.e. automatic generation	Desiderata	Partial implementation.

#	Functionality	Rate	Implementation status (July 2013)
	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)		
13	Keep extensive logs regarding all changes made by the users in the tool	Desiderata	Partial implementation.
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).	Desiderata	Implementation in progress.
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 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).	Desiderata	Postponed for 2014.
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	Postponed
17	Communicate to the centers the info of awarded projects (LDAP) "Project ID, User Accounts, etc."	Essential	Not yet scheduled
18	Create a report of reviewers, with past historical information (reviews attributed and reviews in previous calls), including passwords	Essential	Not yet scheduled
19	Create a report of all persons involved in past and present calls (PIs, collaborators) with history (call, proposal ID, ...)	Essential	Partially implemented
20	Guarantee a highly secure log-in system (highly secure password)	Essential	Partly fulfilled.

Table 6: DECI-PPR-tool complete requirement list

6.3 Questionnaire on Big Data

This survey² aims to gather information on the needs for data storage and data management of users and user communities using the PRACE infrastructure with needs for orders of Terabytes or even Petabytes of data. Users or Communities currently not using the

² Methodologically a pure on-line survey is considered to not provide sufficiently valid results. Thus the survey needs to be accompanied by direct interviews with users and a few representatives of user communities.

infrastructure should answer the questions according to their plans. This information will help to develop a data strategy for PRACE and the HPC centers.

This survey has 16 questions and is divided into five parts, while the second to fourth all deal with data workflow aspects:

- Data Characteristics
- Data Movement
- Data Sharing
- Data Post-Processing
- Other

All questions are related to resources and services which are or should be available in relation with calculations on the PRACE infrastructure. The survey does not discriminate between possible differences for Tier-0³ and Tier-1⁴ systems. In case differences are expected, please indicate that.

To better understand those needs on data, we ask questions about the specification of the data and the workflow. This workflow covers in more detail questions like

- Where does your data come from? (Data Movement)
- What sort of post-processing are you performing? (Post Processing and Data Analysis)
- Where do you store your processed data? (Data Movement)
- Should your processed data be made available and to whom? (Data Sharing)
- Could you describe the storage size and type used for each phase? (Data Characteristics)
- Would you like to enhance your workflow and how could this be achieved? (Other)

Final Remark: If any of the questions sounds too technical to you, just note that and do not answer.

Data Characteristics

- 1 Nature of the data
 - 1.1 How can your data be described (number of files, average size of a file, meta data information, type of data (ASCII or Binary; optionally more details for binary data: images, sound-files, ...)
 - 1.2 How would you estimate the overall data volume of your project?
 - 1.3 How would you estimate I/O volume per a typical processing job?
 - 1.4 How is the distribution and use of scratch data, job intermediate data and result data?
 - 1.5 Do you have needs for the access to structured data (HDF, NetCDF, relational databases, ...)?
 - 1.6 Are you implementing parallel I/O or a specific I/O strategy in your simulation code (pNetCDF, MPI-IO, POSIX approach, dedicated IO program, ...)?
 - 1.7 Do you have requirements for the management of the data, e.g. the handling of metadata and the querying of files?
- 2 Distribution of data into the PRACE infrastructure
 - 2.1 Do you need to store data on multiple PRACE sites?
 - 2.2 Do you expect to reuse data stored on PRACE?

³ Access provided to Tier-0 systems through PRACE-calls: <http://www.prace-ri.eu/Call-Announcements>

⁴ Access provided to Tier-1 systems through DECI-calls: <http://www.prace-ri.eu/DECI-Projects>

- 3 Are there needs for peak storage in PRACE (during a project call)?
 - 3.1 For how long (days/weeks/months)?
 - 3.2 How much data is that?
 - 3.3 Where would you like to have your data stored?
- 4 Are there needs for longer term storage in PRACE (between project calls or after an allocation period)?
 - 4.1 For how long (months or years)?
 - 4.2 How much data is that?
 - 4.3 Where would you like to have your data stored?
- 5 Service Level Description
 - 5.1 What are your reliability requirements?
("safe storage" versus "can be recovered easily by other means")
 - 5.2 What are your availability requirements?
("always online access" versus "archived data")
 - 5.3 Is versioning for the data required?
 - 5.4 Are there any special privacy restrictions required on the data
(Read only/Read-Write access rights depending on user categories, strongly-enforced access rights, data encryption)?

Data Movement

- 6 How much data do you need to import and export from and to PRACE for your calculations?
(volume, frequency)
 - 6.1 Projects in the coming years
 - 6.2 Jobs related to these projects
- 7 Where does your data come from and where do you send your data from PRACE?
 - 7.1 Within the PRACE infrastructure
 - 7.2 From/to your own site or scientific large scale equipment
(telescope, sequencer, accelerator, network of sensors, ...)
 - 7.3 From/to another HPC or data infrastructure
(Eudat, EGI, national/regional centers, XSEDE, ...)
 - 7.4 Do you know how fast your internet connection to the PRACE infrastructure is?
- 8 Tools/protocols required/supported for data transfers:
(if this sounds too technical to you, just don't answer)
 - 8.1 Which tools are supported at your site/infrastructure?
 - 8.2 Which tools are you familiar with and using?
 - 8.3 Which other tools are you interested in or want to use?

Data Sharing

- 9 If you share or need to share your data with other users, groups or communities:
 - 9.1 Is data to be shared among PRACE systems?
 - 9.2 Will data be shared among Projects members?
 - 9.3 Should data be shared with other research infrastructures and which ones?
 - 9.4 What is the size of data to be shared with others partners?
 - 9.5 For how long do you need to share your data with others partners?
 - 9.6 Is public access to the data needed?

- 10 Data sharing technologies:
 - 10.1 What technology/service do you use for sharing?
(if this sounds too technical to you, just don't answer)
 - 10.2 Are you interested in advanced interfaces for sharing your data (e.g. web-portals)?
 - 10.3 Which kind of tool or service would you prefer to use in the future?
 - 10.4 Are there different sharing needs during a project period compared to the time in between project periods (PRACE/DECI calls)?
(e.g. private data that may become publicly available)

Data Post-Processing

- 11 Post-processing and data analysis
 - 11.1 Are you doing or planning to do post-processing using PRACE resources?
 - 11.2 If yes, which tools are you familiar with and using?
 - 11.3 Could you reduce the need of data movement, if you could post-process inside PRACE?
 - 11.4 Do you have needs for remote visualization during or after your jobs?
 - 11.5 Are you investigating on novel data analysis approaches using Map/Reduce or NoSQL?

Other

- 12 Are there sufficient guidelines available on how you should deal with data in PRACE?
- 13 Would you like to have best-practices-guides for reading/writing files efficiently on large scale simulations?
- 14 What problems do you have with data management in PRACE and in general?
- 15 What could be done to enhance your workflow?
- 16 Is there any important question that we have missed?

6.4 Methodology for File Transfer Evaluation

The following sections are a copy of the internal document „*Methodology for File Transfer Evaluation*“. Ever since the formulation document is mentioned, this chapter is meant. The references and all other numbering are adjusted to match this deliverable's list of contents, tables, figures and the references.

6.4.1 Introduction

Objective of the sub-task “*New file transfer technologies*” is to evaluate alternatives to GridFTP, which is currently provided as the only core service for bulk data transfer.

This document aims to define a common methodology for evaluating file transfer technologies that are new for PRACE, i.e. not yet officially supported. There are no specific technologies specified in this document since the methodology has designed to be independent from a specific software solution.

The main reference for this document has been a similar work being carried out by the Energy Sciences Network (ESnet) operated by LLNL and funded by the US DoE⁵.

⁵ <http://fasterdata.es.net>

This activity is being tracked by a dedicated page in the internal PRACE Wiki⁶.

6.4.2 Definitions

The following table fixes some important definitions related to a file transfer activity that will be considered.

Measure	Definition (unit)
Capacity	Link Speed (Gbps)
Narrow Link	Link with the lowest capacity along a path [see Figure 10]
Capacity of the end-to-end path	Capacity of the Narrow Link
Utilized Bandwidth	Current Traffic Load
Available Bandwidth	= (Capacity) – (Utilized Bandwidth)
Tight Link	Link with the least available bandwidth in a path [see Figure 10]
Bandwidth Delay Product (BDP)	The number of bytes in flight to fill the entire path. BDP = (Capacity) * (RTT)

Table 7: File Transfer Measures Definitions

Figure 1 provides an example for determining narrow and tight links of a network path.

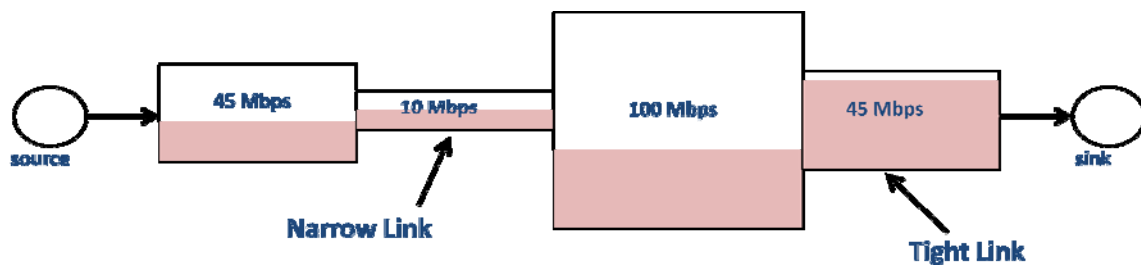


Figure 10: Graphical example for narrow and tight network links

Following the formula stated in Table 7, the BDP for a network with 1Gbps of capacity and 50ms of RTT is:

$$\text{BDP} = 1000\text{Mbps} * 0,05\text{s} = 50\text{Mb} \text{ (6,25MBytes)}$$

6.4.3 Hardware and Configuration Requirements

It is assumed that different persons will be involved in the evaluation of different file transfer tools by using different network paths with unpredictable network conditions.

Defining hardware requirements for the tests is a solution in order to mitigate as much as possible the effect of different conditions. These requirements have been identified and described in the following sections.

TCP Buffer Size

A host system with a GNU/Linux operating system supporting TCP buffer auto-tuning must be used. Auto-tuning technique allows receiver buffer size (and TCP window size) to be dynamically updated for each connection maximizing the action of a congestion algorithm, which is recommended to be “cubic” or “htcp” as documented here⁷.

⁶ <https://prace-wiki.fz-juelich.de/bin/view/Prace2IP/Wp10/Task2/FileTransferTechnoSub-taskActivities>

⁷ <http://fasterdata.es.net/host-tuning/linux/expert/>

Recent versions of Linux (version 2.6.17 and later) support auto-tuning with a default maximum value for the TCP buffer size of 4MByte (4194304 bytes)⁸:

- memory reserved for TCP receiver buffers

```
user@sender_host:~# sysctl net.ipv4.tcp_rmem
net.ipv4.tcp_rmem = 4096 87380 4194304
```
- memory reserved for TCP sender buffers

```
user@sender_host:~# sysctl net.ipv4.tcp_wmem
net.ipv4.tcp_wmem = 4096 16384 4194304
```

It is suggested to increment the maximum value for both sender and receiver buffers, depending from the network card and the BDP measured. The following references help to check whether the maximum TCP buffer size is coherent with the measured BDP. As example, for a host equipped with 10G NIC and RTT delay below 100ms, is preferable to set a value greater than 4MB (16MB or 32MB):

<http://fasterdata.es.net/host-tuning/linux>

<http://www.psc.edu/index.php/networking/641-tcp-tune>

MTU and Jumbo Ethernet Frames

Ethernet's maximum frame size of 1500 bytes is not optimized for Gigabit Ethernet network cards and can actually inhibit the ability of applications to take full advantage of a high network capacity.

This limitation can be overcome by changing the MTU to a value of 9000 allowing Ethernet frames with a payload of 9000 bytes. Assuming `eth0` as the name of the network interface, the MTU can be changed with the following command:

```
user@sender_host:~# ifconfig eth0 mtu 9000
```

Permanent changes take effect by modifying network configuration files, dependently from the specific Linux distribution installed⁹.

Disk performance

Before to run any test, it is absolutely required to check performance of the disks subsystem involved. I/O benchmarks like "*hdparm*", "*bonnie++*" and "*iozone*" could be used to test performance of I/O operations on the disk.

Network capacity

Tests will be executed over both public Internet and private PRACE network.

For public Internet the only requirement is that the user end-point is plugged to a network with the following minimum requirements¹⁰:

- RTT below 70ms
- 0% of packet lost
- Jitter not above 1ms

For hosts connected to the internal PRACE network, no minimum requirements are set.

⁸ To check if the auto-tuning is active, the file "`/proc/sys/net/ipv4/tcp_moderate_rcvbuf`" must be present and with value equal to 1.

⁹ <http://www.cyberciti.biz/faq/centos-rhel-redhat-fedora-debian-linux-mtu-size/>

¹⁰ User-side requirements can be checked with online free tools like <http://pingtest.net/>

Requirements summary

Requirement	Description
TCP Buffer sizing	TCP buffer auto-tuning supported. Maximum Buffer Size adjusted with the BDP.
MTU and Jumbo Frames	Network cards with MTU=9000
Disk performance	I/O performance better than Network performance
Network Capacity for Public Internet	<ul style="list-style-type: none"> - RTT < 70ms - Packet Loss = 0% - Jitter <= 1ms

Table 8: File Transfer Requirements list

6.4.4 Methodology

The proposed methodology must be able to:

- Produce assessments in a consistent manner across different sites and different network paths;
- Consider production conditions and any network turbulence which might occur;
- Assess performance for different types of workloads and different numbers of parallel streams;
- Gather and record results of the evaluation for each technology by using a well defined template;
- Create a straightforward way to qualify and compare results;
- Provide well defined test-cases;

In addition to a quantitative assessment, also factors like reliability, footprint or intrusiveness, maintenance, code maturity, support, should be considered and qualitatively evaluated.

Tests must be executed on both PRACE network and public Internet.

Production Conditions

Before running a test, a report on the network status must be taken. This implies to define at least the Bandwidth Delay Product (BDP), which is calculated multiplying the capacity of the network path (or the narrow link, if any) and the Round-Trip delay Time (RTT):

$$\text{BDP} = (\text{Capacity}) * (\text{RTT})$$

This gives a measure of the network congestion and the ability to compare different file transfer tools under similar values for the BDP.

Data sets

Transferring a large number of small files is significantly different from transferring few large files in terms of performance. Also the directory depth or tree affects performance significantly.

In general, a user should be able to optimize the dataset that has to be transferred, e.g. by using archiving, compression and remote synchronization techniques.

Two dataset are defined to take into account these case studies.

- **Dataset A (Many Small files):**
 - Number of files: ≥ 100

- Size of each file: $\geq 1\text{GB}$
- Directory tree: ≥ 1 level
- **Dataset B (Few Large files)**
 - Number of files: ≤ 10
 - Size of each file: $\geq 100\text{GB}$
 - Directory tree: = 1 level

Workload

There is not a specific study and/or survey figuring out the average amount of data transferred across PRACE sites. Independently from this lack of understanding, it is recommended to test different size of workloads and to study how tools scale. Taking into account the storage availability for this test, three workloads are considered:

- **Workload A: 100GB**
- **Workload B: 500GB**
- **Workload C: 1000GB (1TB)**

Parallel Streams

Only tools that support data transfer parallelism can be considered.

Choosing the number of parallel streams is not a simple task because performance could decrease with high number of streams. It mainly depends from the memory availability at the end points.

Several studies have shown that in practice using between 4 and 8 streams are usually sufficient. 16 streams may be useful only in case of bad performance found with 4 and 8. Above 16 is basically wasting resources.

So it is recommended to run test with 3 different numbers of streams:

- **Parallel Streams Configuration A: 4**
- **Parallel Streams Configuration B: 8**
- **Parallel Streams Configuration C: 16**

Qualitative Factors

It has been considered as valuable to take into account also qualitative factors that are not strictly related to performance of a specific file transfer tool.

Factors like reliability are important for providing a complete feedback whether deciding to include a specific file transfer tool into data services for PRACE.

Evaluation could be provided by using a ranking from 1 (really bad) to 5 (really good) along with a short comment specifying the motivation of the mark.

Recommended factors to be considered are:

- Reliability
- Footprint (Intrusiveness)
- Maintenance
- Fault Tolerance
- Code Maturity
- Community Acceptance

6.4.5 Test cases

Fixed a medium, which could be Internet or the private PRACE network, and taking into account of the methodology above mentioned, there will be **18 runs** to execute for each specific tool. The following table shows an example for two specific dataset types (100 files for Dataset A against 1 file for Dataset B).

#Run	DataSet	Workload	Parallel Streams
1	A (100 files of 1GB)	A (100GB)	A (4)
2	A (100 files of 1GB)	A (100GB)	B (8)
3	A (100 files of 1GB)	A (100GB)	C (16)
4	A (100 files of 5GB)	B (500GB)	A (4)
5	A (100 files of 5GB)	B (500GB)	B (8)
6	A (100 files of 5GB)	B (500GB)	C (16)
7	A (100 files of 10GB)	C (1000GB)	A (4)
8	A (100 files of 10GB)	C (1000GB)	B (8)
9	A (100 files of 10GB)	C (1000GB)	C (16)
10	B (1 file of 100GB)	A (100GB)	A (4)
11	B (1 file of 100GB)	A (100GB)	B (8)
12	B (1 file of 100GB)	A (100GB)	C (16)
13	B (1 file of 500GB)	B (500GB)	A (4)
14	B (1 file of 500GB)	B (500GB)	B (8)
15	B (1 file of 500GB)	B (500GB)	C (16)
16	B (1 file of 1TB)	C (1000GB)	A (4)
17	B (1 file of 1TB)	C (1000GB)	B (8)
18	B (1 file of 1TB)	C (1000GB)	C (16)

Table 9: File Transfer test cases with at least 18 runs each

6.4.6 Template for testing the data transfer tool

Results must be collected by data sheets based on a predefined layout. A data sheet will include quantitative data as well as information about the test bed used. It acts as a data base from which structured information can be further elaborated, e.g. performance with a fixed dataset type and different workloads and parallel streams, performance with a fixed workload and different dataset type and parallel streams, etc...

Information can be presented in table and/or graphic format (recommended).

General Information			
Tool	Site A	Site B	
BBCP	CINES	CEA	
Network			
Type	Capacity	RTT	BDP
Internet	200Mbps	50ms	1250 KBytes
Hosts configuration			

Max TCP Buffer Size (Site A)		Max TCP Buffer Size (Site B)		
net.ipv4.tcp_rmem	net.ipv4.tcp_wmem	net.ipv4.tcp_rmem	net.ipv4.tcp_wmem	
4194304	4194304	4194304	4194304	
Quantitative Assessment				
Run#ID	Dataset Type	Workload	Parallel Streams	Throughput (Mbps)
1	A (100 files)	100GB	4	184.75
2	A (100 files)	100GB	8	192.25
3	A (100 files)	100GB	16	193.10
4	A (100 files)	500GB	4	144.07
5	A (100 files)	500GB	8	121.89
6	A (100 files)	500GB	16	166.27
7	A (100 files)	1000GB	4	184.75
8	A (100 files)	1000GB	8	192.25
9	A (100 files)	1000GB	16	193.10
10	B (1 file)	100GB	4	144.07
11	B (1 file)	100GB	8	121.89
12	B (1 file)	100GB	16	166.27
13	B (1 file)	500GB	4	184.75
14	B (1 file)	500GB	8	192.25
15	B (1 file)	500GB	16	193.10
16	B (1 file)	1000GB	4	144.07
17	B (1 file)	1000GB	8	121.89
18	B (1 file)	1000GB	16	166.27
Qualitative Assessment				
Factor	Rank (1 – 5)	Comment		
Reliability	4	No crashes reported during the tests.		
Footprint Intrusiveness	5	Minimal. It doesn't require administrative rights. Can be installed by a normal user.		
Maintenance	5	No maintenance required by system administrators.		
Fault Tolerance	1	Bad, the tool doesn't provide "restart-after-fail" capabilities.		
Code Maturity	3	Good, first version released in 2011, last version (v2.5) on April 2013.		
Community Acceptance	4	Good. Number of users requesting this tool at both sites is growing.		

Table 10: Example of a filled file transfer evaluation sheet, here for bbcp between CINES and CEA

6.5 Performance Measurement of Remote Visualization

VirtualGL and TurboVNC setup

The tests for RCM (VirtualGL + TurboVNC) have been run on the CINECA infrastructure by opening a session from local TurboVNC client connecting to the remote visualization nodes.

The frame rate has been measured with `tcbench`, a tool included in the VirtualGL bundle that grabs a small portion of the VNC window and counts the number of different frames in a fixed amount of time. In the presence of a running OpenGL application which changes continuously the image, this tool seems to properly evaluate the real frame rate. Two graphics test applications have been used: the very light `vglsphere` app included in VirtualGL and a much heavier Unigine game benchmark. There have been also different TurboVNC settings of image quality and compression type.

Tests have been performed on different network connections and with different platform clients:

- RVN node: GPU: Quadro Fx 1700 CPU core: Xeon E5540 2.5 Ghz
- visual node: GPU: Quadro Fx Tesla M2070 CPU core: Xeon E5645 2.4 Ghz

home urban ADSL italy Windows client	remote fps	lossless zlib	jpeg max	jpeg hi quality	jpeg low quality	jpeg compr 12
vglsphere on visual	190-200	3	2.2	6	12	20
vglsphere on rvn	120-140	3	2.2	6	15	20
UniGine on visual	30-50	0.3	1.5	4	7	19
UniGine on rvn	8-15	0.3	1.2	5	7.5	9

glxspheres 1280x960 on rvn Cineca LAN Windows client:

client compression	server cpu %	client cpu % PLINK	client cpu % vncviewer	server fps	client fps
lossless zlib	35-40%	15%	15%	180	4-5
jpeg max quality (LAN)	95%	37 %	35 %	35-70	28-39
jpeg med quality (default)	90%	30 %	40 %	55-60	57-58
jpeg low quality (WAN)	90%	15%	35%	65-70	64-67

glxspheres 1280x960 on rvn Cineca LAN Linux client:

client compression	server cpu %	client cpu % ssh	client cpu % vncviewer	server fps	client fps
lossless zlib	45-50%	12%	8-15%	150-170	6
jpeg max quality (LAN)	95%	20 %	32 %	35-40	38-39
jpeg med quality (default)	93%	12 %	28 %	55-59	56-58
jpeg low quality (WAN)	92%	8%	24%	62-75	69-72

UniGine 1024x768 on rvn Cineca LAN Windows client:

client compression	server cpu %	client cpu % PLINK	client cpu % vncviewer	server fps	client fps
lossless zlib	50%	15%	7%	8-9	2-4
jpeg max quality (LAN)	60%	18%	15%	7-12	6-9
jpeg med quality (default)	70%	10 %	15 %	7-12	6-9
jpeg low quality (WAN)	60%	10%	12%	7-12	7-9

UniGine 1024x768 on visual Cineca LAN Windows client:

client compression	sshd tunnel login cpu %	server cpu %	client cpu % PLINK	client cpu % vncviewer	server fps	client fps
lossless zlib	8-40%	20%	12%	7%	25-60	2
jpeg max quality (LAN)	20-40%	80%	35%	26%	25-35	20-32
jpeg med quality (default)	25-35%	67%	26 %	29 %	35-50	37-42
jpeg low quality (WAN)	6-10%	65%	17%	22%	40-50	41-42

UniGine 1024x768 on visual Cineca LAN Linux client:

client compression	sshd tunnel	server	client cpu	client cpu	server	client
--------------------	-------------	--------	------------	------------	--------	--------

	login cpu %	cpu %	% ssh	% vncviewer	fps	fps
lossless zlib	11-15%	22-27%	12-16%	8-12%	30-40	2-3
jpeg max quality (LAN)	18-20%	53-76%	32-44%	36-56%	25-40	21-31
jpeg med quality (default)	14-23%	67-72%	24-28%	36-48%	28-37	32-35
jpeg low quality (WAN)	3-7%	35-43%	12-16%	28-32%	35-45	31-36

UniGene 1024x768 on visual low speed ADSL Windows client:

client compression	sshd tunnel login cpu %	server cpu %	client cpu % PLINK	client cpu % vncviewer	server fps	client fps
jpeg med quality (default)	1-3%	5%	3%	2%	40-50	1-2
jpeg low quality (WAN)	2%	7%	2%	7%	39-50	2-3
jpeg ultra low quality (12%)	1-3%	10%	2%	7%	40-60	5-6

NICE DCV setup

UniGene 1024x768 on visual low speed ADSL Windows client:

client compression	client cpu % vncviewer	server fps	client fps
default settings (low quality)	10%	40-70	4-6

UniGene 1024x768 on visual Cineca LAN Windows client:

client compression	client cpu % vncviewer	server fps	client fps
default settings (adaptive)	28%	40-70	19-24

6.6 iRODS Evaluation Forms

The following sub-sections details the evaluations results of the different iRODS functions. The seven sections are originally seven individual documents. The references therein refer to the References section of the respective document/sub-section.

6.6.1 iRODS Workflow-Objects Evaluation

Introduction

The goal of this document is to evaluate the workflow objects feature. Workflow objects are a new feature introduced in iRODS 3.2

Purpose

The common definition of a workflow is that it allows to chain and control tasks in order to perform a complex processing. It consists of a sequence of steps which execute a series of computation or data manipulation. Each operation in the step may use data provided as input to the workflow or data created in a previous steps and may produce output data.

In the iRODS context, the operations in a workflow objects can be mapped to micro-services, so that a complete workflow can be seen as an iRODS rule that will be executed at the server side. The implementation of a workflow is depicted as an iRODS active object where both a workflow file and a set of related parameter files are ingested into iRODS. A WSO (Workflow Structured Object) is associated to each workflow file for providing an iRODS

collection-type environment for running the workflow. The WSO will gather parameters files needed to run the workflow as well as input files needed for the workflow execution. A set of run directories will be attached to this structure that houses the results of executions. The WSO is created as a mount point in the iRODS logical collection hierarchy.

A parameter file contains information needed for executing the workflow as well as information about files that need to be staged in before the execution or staged out after the execution.

When a parameter file is ingested into a WSO, a run file is automatically created which will be used to execute the parameter file with the associated workflow. When a workflow execution occurs, a run directory is created automatically for storing the results of this run.

Unlike other software, the iRODS implementation of workflows doesn't provide a graphical interface used to model the various steps of the workflow. Rather, it is based on the standard iRODS rules and micro-services. Additionally, it manages automatically the execution within the iRODS environment, preventing the end-user to ingest or retrieve from/into iRODS, files needed for the execution or produced by the workflow.

The evaluation consists in building workflows using different set of parameters for checking both the workflow environment creation and execution.

The workflow object feature is intended to be used by end-users. However, in order to run this feature end-users need to be defined as "rodsadmin" to allow collections to be mounted.

Responsibility

SITE	ROLE/TASK
IDRIS	Full evaluation

References

- [1] https://www.irods.org/index.php/Release_Notes_3.2
- [2] https://www.irods.org/index.php/Introduction_to_Workflow_as_Objects
- [3] https://www.irods.org/index.php/Workflow_Objects_%28WSO%29
- [4] https://www.irods.org/index.php/Realized_Objects
- [5] <https://www.irods.org/index.php/glossary>
- [6] <https://www.irods.org/index.php/Downloads>
- [7] <https://bscw.zam.kfa-juelich.de/bscw/bscw.cgi/919814>
- [8] <https://bscw.zam.kfa-juelich.de/bscw/bscw.cgi/819591>

Tested Components

The workflow object evaluations focus on the software. The related documentation is specified at [1], [2], [3], [4].

The iRODS release used for the evaluation is 3.2. It can be downloaded at [6]

The evaluation requires applying a patch related to the file "msssoStructFileDriver.c" (unnumbered for now). This patch has been delivered by the iRODS developers and it has to be installed and iRODS has to be recompiled before testing. It can be downloaded at [7].

iRODS rule files will be used for testing the workflow execution and random data test files provided at [8] will be used for testing the stage in/out and copyout functionalities.

Tested Features

The workflow objects evaluation will focus on 3 major points:

- a) The set up of the iRODS workflow environment:

The tests will consist of evaluating how the iRODS workflow environment is built before the workflow execution. The set of commands to build the environment will be run, checking for the real impact.

- b) The workflow execution:

The test will consist of executing a workflow and checking how the workflow environment is modified. This test will be performed with a basic parameter file.

- c) The various parameters file settings:

The tests of the various parameters described in the workflow parameter file will be evaluated:

- INPARAM: describe a input parameter
- FILEPARAM: identify files that are used as input parameters (INPARAM). It is needed to stage back outputs.
- STAGEAREA: identify the stage area where the workflow execution is performed
- STAGEIN: stage in files from anywhere in iRODS to the stage area
- STAGEOUT: move files from the stage area to the iRODS WSO
- COPYOUT: leave a copy in the stage area and make a copy in the iRODS WSO (useful if it is needed for subsequent workflow execution)
- NOVERSION: turn off the versioning of results
- CLEANOUT: clear the stage area after execution
- CHECKFORCHANGE: check is the file being checked has changed since the previous execution of the workflow. If the file has been changed then the workflow is executed otherwise it is not.

Non Tested Features

Rules and micro-services are used by the workflows but will not be tested as such.

Test Phases

The test phases have been described in paragraph 6. The execution order is a); b); c)

Tests Environment

SITE	TEST ENVIRONMENT
IDRIS	Fedora release 17 (Beefy Miracle) IBM System x3655 4 Proc Dual-Core AMD Opteron(tm) Processor 2218 2 GB RAM 1Gb Ethernet card

Testing Methodology

The test will focus on the functional aspect. So, we will check whether the functionality is provided that if whether it does what is supposed to do, the easiness of use, the reliability and maintainability.

Tests Description

Test of the workflow environment set up

Initial conditions	<ul style="list-style-type: none"> irods user account, defined as “rodsadmin”
Parameters and input data	<ul style="list-style-type: none"> the “testw.mss” workflow file that describes the workflow a basic parameter file “testw.mpf”
Test procedure	<ul style="list-style-type: none"> Create an iRODS collection and ingest the workflow file (data type msso) <pre>imkdir /IDRIS/home/pr1f02is/workflow</pre> <pre>iput -D "msso file" ./testw.mss /IDRIS/home/pr1f02is/workflow/testw.mss</pre> <ul style="list-style-type: none"> Create a collection and mount that collection as a Workflow Structure Object associated with the workflow file testw.mss <pre>imkdir /IDRIS/home/pr1f02is/workflow/testWF</pre> <pre>imcoll -m msso /IDRIS/home/pr1f02is/workflow/testw.mss /IDRIS/home/pr1f02is/workflow/testWF</pre> <ul style="list-style-type: none"> Ingest a parameter file (testw.mpf) in the WSO collection testWF. <pre>iput testw.mpf /IDRIS/home/pr1f02is/workflow/testWF</pre>
Expected result	a “*.run” file is automatically created in the WSO environment that will be used for the workflow execution later on
Test result	<pre>ils -l</pre> <pre>/IDRIS/home/pr1f02is/workflow/testWF:</pre> <pre>pr1f02is mssoSt demoResc 2392 2013-04-</pre> <pre>23.15:19 & testw.mpf</pre> <pre>pr1f02is mssoSt demoResc 33554412 2013-04-</pre> <pre>23.15:19 & testw.run</pre>

	The test result matches the expected result.
--	---

Test of the workflow execution

Initial conditions	
Parameters and input data	<ul style="list-style-type: none"> the “<i>testw..run</i>” file
Test procedure	<ul style="list-style-type: none"> Launch the workflow execution using the *.run file: iget testw.run -
Expected result	The workflow is executed
Test result	<p>Command result is</p> <pre>>>> ecrifich/info: File=/home/iRODS/Vault/home/pr1f02is/workflow/ testw.mss.cacheDir0/testw.runDir/fichin-1 NBlock=128 BlkSize=512 Workflow ecrfich Executed Successfully at 2013-4-23 15h:14m:12s:</pre> <p>The test result matches the expected result.</p>

Test of the parameter file settings

Note:

- Thereafter, a new test is performed to evaluate the setting of a given parameter.
- A new parameter file is created and ingested into iRODS, thus producing each time a new *.run file. Indeed, testing the different parameter settings on the same file introduces some edge effect that we didn't want to face.
- For each setting, the *test procedure* and the *expected results* are described in the same way so we will detail only the *parameters and input data*, the *initial conditions* and the *test result*.

Initial conditions	<ul style="list-style-type: none"> STAGEAREA parameter set
Parameters and input data	<ul style="list-style-type: none"> the “X.run” file
Test procedure	<ul style="list-style-type: none"> Launch the workflow execution using the *.run file: iget X.run -
Expected result	The workflow is executed using the parameters describe in the parameter file
Test result	In the current release, the STAGEAREA parameter only refers to the “bin” directory of the iRODS server. Changing

	<p>this value introduces a wrong behaviour.</p> <p>The test result doesn't match the expected result.</p> <p>Note:</p> <p>The stagearea is defined on the local machine.</p>
--	---

Initial conditions	<ul style="list-style-type: none"> ▪ STAGEIN parameter set
Parameters and input data	<ul style="list-style-type: none"> ▪ the file to stage in which can be in the WSO environment or anywhere in iRODS
Test result	<p>The stage in action on a file, copies the file in the bin/cmd directory of the iRODS server. This directory is not the STAGEAREA (even fixed to "bin" for now), so should be copied in "bin". The file owner becomes the iRODS admin.</p> <p>The copy is performed properly for a file in the WSO or anywhere in the iRODS environment.</p> <p>The test result doesn't match the expected result.</p>

Initial conditions	<ul style="list-style-type: none"> ▪ INPARAM-FILEPARAM parameters set
Parameters and input data	<ul style="list-style-type: none"> ▪ INPARAM *Arg="fichin-1" ▪ FILEPARAM *Arg
Test result	<pre>ils -l /IDRIS/home/pr1f02is/workflow/testWF/testw.runDir: pr1f02is msoSt demoResc 741 2013-04- 23.15:22 & stdout pr1f02is msoSt demoResc 262152 2013-04- 23.15:22 & fichin-1</pre> <p>The test result matches the expected result.</p> <p>Note:</p> <p>If the parameter FILEPARAM is not set, the "fichin-1" file will remain in the stagearea (bin) and will not be copied back in the WSO.</p> <p>The parameter STAGEOUT fichin-1 has no impact/effect in this case.</p>

Initial conditions	<ul style="list-style-type: none"> ▪ STAGEOUT parameter set
Parameters and input data	<ul style="list-style-type: none"> ▪ the file to stage out from the stagearea to the WSO
Test result	<p>The file is staged out as expected.</p>

	<p>The test result matches the expected result.</p> <p>Note:</p> <p>There is currently only one stagearea available to the users (bin). We found here a security issue as any file can be staged out from this directory by any user (all files are owned by the iRODS administrator).</p> <p>Please note that no error message appears if the file that has to be staged out doesn't exist for any reason so cannot be staged out.</p>
--	--

Initial conditions	<ul style="list-style-type: none"> ▪ CLEANOUT parameter set
Parameters and input data	<ul style="list-style-type: none"> ▪ None
Test result	<p>The files which have been staged in remain in the bin/cmd directory and the files defined as INPARAM remain in the stagearea.</p> <p>The test result doesn't match the expected result.</p>

Initial conditions	<ul style="list-style-type: none"> ▪ NOVERSION parameter set
Parameters and input data	<ul style="list-style-type: none"> ▪ None
Test result	<p>The NOVERSION parameter stops the versioning of the execution directories. During our test, the versioning still goes on.</p> <p>The test result doesn't match the expected result.</p>

Initial conditions	<ul style="list-style-type: none"> ▪ CHECKFORCHANGE parameter set
Parameters and input data	<ul style="list-style-type: none"> ▪ the file to check which can be in the WSO environment or anywhere in iRODS
Test result	<p>Files are not stage in/out and the new execution directory is not created. Nevertheless, the rules and micro-services are executed.</p> <p>The test result doesn't match the expected result.</p>

Conclusions

The workflow objects feature provides some interesting mechanisms to help the users to run iRODS workflows in an integrated environment.

The parameter file allows running workflows in a different context and provides a way for the user to easily interact with the iRODS environment.

We found in this first release, several dysfunctions as well as a security issue. These issues are going to be reported to the iRODS developers.

However, this feature remains difficult to use and to integrate in the user development environment, so that users can hesitate to use it.

6.6.2 iRODS PAM-LDAP-Authentication-Evaluation

Introduction

This document is within the scope of the iRODS sub-task in WP10/Task T10.2 “Evaluating data services”. It provides a homogeneous way to evaluate features and software for a better coherency within the iRODS working group. This document focuses on a new feature in iRODS 3.2: PAM/LDAP Authentication.

Purpose

PAM provides a mechanism for applications to detach the task of authentication from the application itself. Most of the time, organizations already have their AA infrastructure and it is a common case where LDAP is included in such a setup. A PAM/LDAP module therefore could be used to authenticate existing LDAP users.

In the iRODS context, users may authenticate themselves using their LDAP password instead of having a separate password in iRODS.

For the sake of security the password exchange is protected (SSL is being used) and then an iRODS-generated password is used which is valid for two weeks by default.

These so called PAM-derived Passwords may be removed by an administrator for specific users.

Evaluation will focus on the setup, the ease of use and security aspects of the feature.

User base for this feature includes end-users as well as administrators.

Responsability

SITE	ROLE/TASK
NIIFI	Full evaluation

References

- [1] https://www.irods.org/index.php/Release_Notes_3.2
- [2] https://www.irods.org/index.php/PAM_Authentication
- [3] https://www.irods.org/index.php/PAM/LDAP_Authentication/Authorization
- [4] https://www.irods.org/index.php/PAM_SSL_Setup
- [5] <https://www.irods.org/index.php/glossary>
- [6] <https://www.irods.org/index.php/Downloads>

Tested Components

The PAM/LDAP Authentication evaluation focus on the software. The related documentation is specified at [1], [2], [3], [4].

The iRODS release used for the evaluation is 3.2. It can be downloaded at [6].

Tested Features

Evaluation will focus on PAM/LDAP integration.

Non Tested Features

Other features of iRODS are not part of this evaluation.

Tests Phases

There are two phases of evaluation:

- a) Check if PAM works
\$IRODS_HOME/server/bin/PamAuthCheck will be used for this test.
- b) Test cases
These tests include iRODS icommands and other iRODS-related activities.

Tests Environment

SITE	TEST ENVIRONMENT
NIIFI – iRODS test server	Cloud VM Debian 6.0.7 2 cores 0.5 vcpu (i.e. 50% guaranteed cpu time) 2GB RAM 1Gbit Ethernet
NIIFI – LDAP server	Cloud VM Debian 6.0.7 1 core 1 vcpu (i.e. 100% guaranteed cpu time) 1GB RAM 1Gbit Ethernet Software: OpenLDAP

PAM configuration (/etc/pam.d/irods) for irods service:

auth	required	pam_ldap.so
------	----------	-------------

Testing Methodology

The test will focus mainly on the functional aspect. It will be checked whether the functionality is provided and it does what is supposed to do. Also the ease of use, reliability, maintainability and security will be evaluated.

Tests Description

Basic PAM test

Initial conditions	<ul style="list-style-type: none"> • niifitest user account in iRODS, defined as "rodsuser" • niifitest user in LDAP • niifitest has password in LDAP only
Parameters and input data	<ul style="list-style-type: none"> • username, that is "niifitest" • PAM asks for the password
Test procedure	<ul style="list-style-type: none"> • cd \$IRODS_HOME/server/bin • ./PamAuthCheck niifitest
Expected result	PamAuthCheck waits for password. Then, it tells whether the user authenticated successfully or not.
Test result	<pre>\$./PamAuthCheck niifitest wrongpassword Not Authenticated \$ \$./PamAuthCheck niifitest ***** Authenticated</pre> <p>The test result matches the expected result.</p>

Detailed tests – PAM-derived password does not exist – 1

Initial conditions	<ul style="list-style-type: none"> • see Basic PAM test • PAM-derived password does not exist • additionally a custom iCAT query (created by admin user) will be used to check expiry timestamp (i.e. lifetime) as well as creation/modification timestamps of PAM-derived passwords on server side: \$ iadmin asq 'select pass_expiry_ts, R_USER_PASSWORD.create_ts, R_USER_PASSWORD.modify_ts from R_USER_PASSWORD, R_USER_MAIN where user_name=? and zone_name=? and R_USER_MAIN.user_id = R_USER_PASSWORD.user_id' ShowPPTimes
Parameters and input data	<ul style="list-style-type: none"> • username, that is "niifitest" • zone name, that is "tempZone"
Test procedure	<p>N: niifitest, A: admin</p> <ul style="list-style-type: none"> • N <ul style="list-style-type: none"> ◦ ls -a .irods/ • A <ul style="list-style-type: none"> ◦ iquest --sql ShowPPTimes niifitest tempZone • N <ul style="list-style-type: none"> ◦ ils

Expected result	There is no .irodsA file present on client side. There is no PAM-derived password present in iCAT. The ils command should not work. It is expected that iRODS will prompt for iRODS password and then fail because this user does not have an iRODS password.
Test result	<pre>\$ ls -a .irods/irodsEnv \$ iquest --sql ShowPPtimes niifitest tempZone No rows found \$ ils Enter your current iRODS password: rcAuthResponse failed with error -826000 CAT_INVALID_AUTHENTICATION</pre> <p>The test result matches the expected result.</p>

Detailed tests – PAM-derived password does not exist – 2

Initial conditions	<ul style="list-style-type: none"> see <i>PAM-derived password does not exist – 1</i>
Parameters and input data	<ul style="list-style-type: none"> username, that is "niifitest" zone name, that is "tempZone" PAM asks for the password
Test procedure	<p>N: niifitest, A: admin</p> <ul style="list-style-type: none"> N <ul style="list-style-type: none"> iinit (type invalid password) ls -a .irods/ A <ul style="list-style-type: none"> iquest --sql ShowPPtimes niifitest tempZone
Expected result	The iinit command should prompt for PAM password. It should fail upon entering an invalid password. As the user failed to authenticate a PAM-derived password should not be present (neither on client side nor in iCAT).
Test result	<pre>\$ iinit Enter your current PAM (system) password: rcPamAuthRequest failed with error -993000 PAM_AUTH_PASSWORD_FAILED \$ ls -a .irods/irodsEnv \$ iquest --sql ShowPPtimes niifitest tempZone No rows found</pre> <p>The test result matches the expected result.</p>

Detailed tests – PAM-derived password does not exist – 3

Initial conditions	<ul style="list-style-type: none"> • see <i>PAM-derived password does not exist – 1</i>
Parameters and input data	<ul style="list-style-type: none"> • username, that is "niifitest" • zone name, that is "tempZone" • PAM asks for the password
Test procedure	<p>N: niifitest, A: admin</p> <ul style="list-style-type: none"> • N <ul style="list-style-type: none"> ◦ iinit (type valid password) ◦ ls -a .irods/ • A <ul style="list-style-type: none"> ◦ iquest --sql ShowPPtimes niifitest tempZone
Expected result	<p>The iinit command should prompt for PAM password. It should succeed upon entering the valid password. After successful authentication a PAM-derived password should be present (both on client side and in iCAT).</p>
Test result	<pre>\$ iinit Enter your current PAM (system) password: \$ ls -a .irods/irodsA .irodsEnv \$ iquest --sql ShowPPtimes niifitest tempZone 1209600 01372080475 01372080475 ShowPPtimes shows that expiry timestamp is two weeks (1209600 seconds). As the PAM-derived password is just created, creation and modification timestamps are the same. The test result matches the expected result.</pre>

Detailed tests – A valid PAM-derived password does exist – 1

Initial conditions	<ul style="list-style-type: none"> • see Basic PAM test • PAM-derived password exists both on client side and in iCAT
Parameters and input data	.irodsA on client side
Test procedure	ils
Expected result	With a valid PAM derived password ils should succeed.
Test result	<pre>\$ ils /tempZone/home/niifitest: The test result matches the expected result.</pre>

Detailed tests – A valid PAM-derived password does exist – 2

Initial conditions	<ul style="list-style-type: none"> • see <i>PAM-derived password does exist – 1</i> • ShowPPTimes will be used
Parameters and input data	<ul style="list-style-type: none"> • username, that is "niifitest" • zone name, that is "tempZone" • PAM asks for the password
Test procedure	<p>N: niifitest, A: admin</p> <ul style="list-style-type: none"> • A <ul style="list-style-type: none"> ◦ iadmin rpp niifitest ◦ iquest --sql ShowPPTimes niifitest tempZone • N <ul style="list-style-type: none"> ◦ ils ◦ cp -p .irods/.irodsA ./oldauth ◦ iinit ◦ diff -q ./oldauth .irods/.irodsA ◦ ils • A <ul style="list-style-type: none"> ◦ iquest --sql ShowPPTimes niifitest tempZone
Expected result	<p>If an administrator issues an 'iadmin rpp' (remove PAM-derived Password) command for the user then icommands should not work until re-authentication (iinit).</p> <p>The ils command should prompt for iRODS password. It should fail because there is no iRODS password for user. The iinit command should prompt for PAM password and succeed if the password is valid. On success a new PAM-derived password is generated.</p> <p>On client side the new .irodsA differs from the backup. The ils command should work now.</p>
Test result	<pre> \$ iadmin rpp niifitest \$ iquest --sql ShowPPTimes niifitest tempZone No rows found \$ ils rcAuthResponse failed with error -827000 CAT_INVALID_USER \$ cp -p .irods/.irodsA ./oldauth \$ iinit Enter your current PAM (system) password: \$ diff -q ./oldauth .irods/.irodsA Files ./oldauth and .irods/.irodsA differ \$ ils /tempZone/home/niifitest: \$ iquest --sql ShowPPTimes niifitest tempZone 1209600 01372144219 01372144219 </pre> <p>Results match expected results, except that ils fails immediately instead of prompting for an iRODS</p>

	password beforehand.
--	----------------------

Detailed tests – A valid PAM-derived password does exist – 3

Initial conditions	<ul style="list-style-type: none"> see <i>PAM-derived password does exist – 1</i>
Parameters and input data	<ul style="list-style-type: none"> username, that is "niifitest" zone name, that is "tempZone"
Test procedure	<ul style="list-style-type: none"> cp -p .irods/.irodsA ./ iexit full ls -a .irods ils cp -p ./irodsA .irods/ ils
Expected result	<p>If the user issues an 'iexit full' command then the PAM derived password is removed from .irods directory on client side.</p> <p>It is expected that ils should prompt for iRODS password and then fail because this user does not have an iRODS password.</p> <p>It should work though after a backup of .irodsA is copied back.</p>
Test result	<pre>\$ cp -p .irods/.irodsA ./ \$ iexit full \$ ls -a .irodsirodsEnv \$ ils Enter your current iRODS password: rcAuthResponse failed with error -826000 CAT_INVALID_AUTHENTICATION \$ cp -p ./irodsA .irods/ \$ ls -a .irods/irodsA .irodsEnv \$ ils /tempZone/home/niifitest: Results match expected results.</pre>

Detailed tests – A valid PAM-derived password does exist – 4

Initial conditions	<ul style="list-style-type: none"> see <i>PAM-derived password does exist – 1</i> ShowPPTimes will be used
Parameters and input data	<ul style="list-style-type: none"> username, that is "niifitest" zone name, that is "tempZone" PAM asks for the password
Test procedure	<p>N: niifitest, A: admin</p> <ul style="list-style-type: none"> A

	<ul style="list-style-type: none"> ◦ iquest --sql ShowPPtimes niifitest tempZone • N ◦ cp -p .irods/.irodsA ./ ◦ iinit ◦ diff -q ./irodsA .irods/.irodsA • A ◦ iquest --sql ShowPPtimes niifitest tempZone
Expected result	<p>Issuing an iinit command extends the lifetime of a PAM derived password.</p> <p>On client side a new .irodsA should be created which differs from the old one.</p> <p>Modification timestamp should be updated in iCAT.</p>
Test result	<pre>\$ iquest --sql ShowPPtimes niifitest tempZone 1209600 01372144219 01372144219 \$ cp -p .irods/.irodsA ./ \$ iinit Enter your current PAM (system) password: \$ diff -q ./irodsA .irods/.irodsA Files ./irodsA and .irods/.irodsA differ \$ iquest --sql ShowPPtimes niifitest tempZone 1209600 01372144219 01372144346</pre> <p>The test result matches the expected result.</p>

Detailed tests – An existing PAM derived password is invalidated on client side – 1

Initial conditions	<ul style="list-style-type: none"> • see <i>PAM-derived password does exist – 1</i>
Parameters and input data	<ul style="list-style-type: none"> • .irodsA at client side
Test procedure	<ul style="list-style-type: none"> • touch -m -t 20131231 .irods/.irodsA • ils
Expected result	<p>The ils icommands should not work with an invalidated PAM derived password.</p> <p>It is expected that iRODS will prompt for an iRODS password and fail because the user has no iRODS password.</p>
Test result	<pre>\$ ils Enter your current iRODS password: rcAuthResponse failed with error -826000 CAT_INVALID_AUTHENTICATION</pre>

	The test result matches the expected result.
--	---

Detailed tests – An existing PAM derived password is invalidated on client side – 2

Initial conditions	<ul style="list-style-type: none"> • see <i>PAM-derived password does exist – 1</i> • .irodsA is invalidated (timestamp changed)
Parameters and input data	<ul style="list-style-type: none"> • .irodsA at client side
Test procedure	<ul style="list-style-type: none"> • touch -m -t <original timestamp> .irods/.irodsA • ils
Expected result	The ils command should not work again if the invalidated PAM derived password's timestamp is changed back to the original value.
Test result	<p>After touch command: \$ ils Enter your current iRODS password: rcAuthResponse failed with error -826000 CAT_INVALID_AUTHENTICATION</p> <p>The test result matches the expected result.</p>

Detailed tests – An existing PAM derived password is invalidated on client side – 3

Initial conditions	<ul style="list-style-type: none"> • see <i>PAM-derived password does exist – 2</i>
Parameters and input data	<ul style="list-style-type: none"> • PAM asks for the password
Test procedure	<ul style="list-style-type: none"> • ils • iinit • ils
Expected result	The ils command should work again after re-authentication (iinit).
Test result	<p>\$ ils Enter your current iRODS password: rcAuthResponse failed with error -826000 CAT_INVALID_AUTHENTICATION \$ iinit Enter your current PAM (system) password: \$ ils /tempZone/home/niifitest:</p> <p>The test result matches the expected result.</p>

Detailed tests – An existing PAM derived password is invalidated on client side – 4

Initial conditions	<ul style="list-style-type: none"> • see <i>PAM-derived password does exist – 2</i> • a backup of the original .irodsA is available
Parameters and input data	<ul style="list-style-type: none"> • .irodsA and a backup of original (valid) .irodsA at client side
Test procedure	<ul style="list-style-type: none"> • ils • cp -p ../irodsA .irods/ • ils
Expected result	The ils command should work again if a backup of the original .irodsA is copied back to .irods directory.
Test result	<pre>\$ ils Enter your current iRODS password: rcAuthResponse failed with error -826000 CAT_INVALID_AUTHENTICATION \$ cp -p ../irodsA .irods/ \$ ils /tempZone/home/niifitest: The test result matches the expected result.</pre>

Conclusions

The PAM/LDAP Authentication feature allows using already existing LDAP Authentication instead of the usual iRODS password authentication. Password exchange is protected (SSL is being used) and subsequent to that an iRODS-generated short term (two weeks) password is used (for other i-commands).

As the PAM-derived password is stored (in a scrambled form) in .irodsA file on the client side, care must be taken to protect this file to assure that an impersonation attack cannot be made. Although there is a way ('iadmin rpp') for the iRODS admin to remove the PAM-derived password for a user, it is still advisable for clients to remove the .irodsA file (e.g. by issuing 'iexit full') when it is not needed and to keep away from making backup copies of it.

As a side note: PAM can be configured to interact with various authentication systems so iRODS could be integrated with those as well. However, integration with other systems is outside of the scope of this evaluation.

6.6.3 iRODS-Ticket-Based-Access-Evaluation

Introduction

The goal of this document is to evaluate the Ticket Based Authentication functionality feature. This is a new feature, first introduced in iRODS 3.1.

Purpose

The tested feature should allow end-users to share data with other people for a limited amount of time or for a limited number of times or till the permission is revoked.

After receiving a ticket (i.e. a string), it should be possible to exploit it as an authenticated user as well as an anonymous user, if such users exists on the server.

For example, it is possible to create a ticket in order to give read access to a given iRODS collection two times for the following two days. The ticket could, after its creation, be sent to a user who, in the next a couple of days, could use it two times to read the content of the given iRODS collection. If the user has no account on the iRODS server and if the anonymous user has been created, the user should be able to authenticate to iRODS as the anonymous user and access to the data described in the ticket as every other user.

Responsibility

The test activity is carried on by CINECA.

References

CINECA evaluated the [ticket based access](#) for irods.

[1] https://www.irods.org/index.php/Ticket-based_Access

See also [iticket](#).

[2] <https://www.irods.org/index.php/iticket>

and:

[3] <https://groups.google.com/d/topic/irod-chat/K3Cbyq0C8nY/discussion>

[4] <svn://irodssvn.ucsd.edu/iRODS/clients/icommands/scripts/>

Tested Components

We tested the iticket functionality of iRODS, introduced first in iRODS 3.1.

We tested on an iCAT (mySQL) enabled iRODS 3.2 server.

Tested Features

Ticket based access.

Non Tested Features

Most iRODS functionality is irrelevant for this activity.

Tests Phases

The test consisted in creating tickets with various parameters values (such as different validity in time and file size) and verify their functionality.

Tests Environment

SITE	TEST ENVIRONMENT
CINECA	DELL <ul style="list-style-type: none"> • 16 x Intel E5530 @ 2.4 GHz • 64 GB • 1 Gb Ethernet card • Debian GNU/linux 6.0 Local FS : GPFS-NFS @ ~160 MB/s R/W iRODS : 3.2

Testing Methodology

We evaluated functionality, easyness of use and reliability.

The icommands have been configured for three different users:

- the owner of the file (password authenticated user)
- a second, different, password authenticated user
- an anonymous user

Tests Description

The tested functionality (*iticket*, *iget*, *iput*) worked as expected for files and directories, resulted easy to use and reliable, but a bug has been found with the subdirectories transfer (see [3], iROD-Chat:9990): the recursive transfer of subdirectory does not work.

This was preventing the usage of this functionality for a production environment. CINECA developed a set of wrapper around the icommands (*bash* scripts) to produce a workaround.

This workaround is now on iRODS SVN [4].

Conclusions

The ticket based authentication proved to be a very useful feature, easy to use and reliable. Even if it had some initial problem, it is now possible to use it as expected.

No particular security concern is foreseen.

6.6.4 iRODS FUSE-Evaluation

Introduction

The goal of this document is to evaluate the FUSE feature. FUSE was introduced in iRODS 1.0.

This feature works for the iRODS client: it adds the *irodsFs* command to the icommands.

Purpose

The tested feature should allow end-users to mount their iRODS home directory on each machine where the feature is installed.

In order to enable the functionality, it is necessary to recompile the icommands (no package is available yes) after having enabled the functionality in the iRODS configuration file and having installed the required libraries. In particular, the FUSE package has to be installed and configured in order to give the users the right to use it (inclusion in fuse group in */etc/groups*).

Once the installation is completed, a user should be able to, for example, mount its iRODS home collection in a mount point of its UNIX workstation and that data via traditional UNIX command line tools.

Responsibility

The test activity is carried on by CINECA.

References

CINECA evaluated the [FUSE](#) support for irods.

[1] https://www.irods.org/index.php/iRODS_FUSE

See also [imcoll](#).

[2] <https://www.irods.org/index.php/imcoll>

and

[3] <https://groups.google.com/forum/#!msg/irod-chat/eL1lQ5z6ot4/somesjc-CQsJ>

Tested Components

We tested the FUSE functionality of iRODS, introduced first in iRODS 1.0.

We tested on an iCAT (mySQL) enabled iRODS 3.2 servers with icommands of the same version.

Tested Features

FUSE (irodsFS) access.

Non Tested Features

Most iRODS functionality is irrelevant for this activity.

Tests Phases

The tests consisted in mounting an iRODS home directory with irodsFS and testing its reliability with I/O stress test.

Tests Environment

SITE	TEST ENVIRONMENT
CINECA	DELL <ul style="list-style-type: none"> • 16 x Intel E5530 @ 2.4 GHz • 64 GB • 1 Gb Ethernet card • Debian GNU/linux 6.0 Local FS : GPFS-NFS @ ~160 MB/s R/W iRODS : 3.2

Testing Methodology

We evaluated functionality, easiness of use and reliability.

The irodsFS command has been used to access data in three ways:

- UNIX command line interface: cp, ls, mv and rsync
- UNIX product account: apache -> apache-user -> mount-point owned by apache-user
- iRODS icommands

Tests Description

The tested functionality worked quite well: the user mounting iRODS home via FUSE was able to read and write data to the mounted directory from the command line with traditional unix commands in the usual way. The same holds true if the access is performed by a product account such as the one running a web server.

A couple of limitations have been revealed:

- it is not possible to use icommands (this is by design, documented on the official web page of the feature);
- it was unstable with rsync (stable elsewhere).

CINECA tested a new patch from Hao Xu (see iROD-Chat:9650 [3]) to resolve the second issue: the problem with the use of rsync is now solved.

Anyway, the first problem (unavailability of icommands) prevents the usage of this functionality for some kind of production environment where data should be writable also by iRODS.

Conclusions

The FUSE module is working well, even if it is not easy to install. It is used in production environment around the world and proved to be quite useful, even if not absolutely reliable: it could be sometimes necessary to umount and remount the collection because sometime the mount process freezes.

6.6.5 iRODS Performance Evaluation

Introduction

The goal of this document is to evaluate the file transfer performance aspect using the iRODS tool, in a test environment. Only the throughput aspect of the performance is studied.

Purpose

As described in [2], iRODS is a data grid software system providing access to storage distributed on multiple sites and heterogeneous hardware and software storage.

Several aspects of this tool could have been studied but this report focus on the performance part.

PRACE already offers two services [5] to users to transfer data across the infrastructure. The first one is GridFTP which is a data transfer tool defined as a core service and the second one is GPFS-MC which is a distributed filesystem (defined as optional). The purpose of this study is to evaluate the iRODS protocol which comes with the iRODS data management tool.

Performance analysis based on various tools has been achieved in the “New File Transfer Tool” task. iRODS wasn’t evaluated in this scope, as it is not a file transfer tool only and provides also a rich and large additional number of data management functionalities far over a simple file transfer tool. Therefore, it is studied as a separate tool in the “iRODS” task.

The major user concern when using a file transfer service is the data access time when data is not located on the site where the user wants to use them. So, the performance to access user data is an important question.

PRACE is a high end HPC infrastructure in Europe. The data used on this infrastructure is in the same order as the compute power it provides. We are talking here about Terabytes of data manipulated across the infrastructure. At this scale, the number of files cannot be the most important element comparing to the volume although it can be also an issue.

iRODS can be setup to allow a cross access from several sites, providing themselves their own iRODS server. In this case, servers are interconnected thru “remote zones” where accesses are restricted to authorized users.

The evaluation consists in transferring files using different set of parameters for evaluating the transfer bandwidth obtained from the end-user point of view.

Note that this evaluation relies on a testbed far from a production environment. It is based on heterogeneous hardware at each site, so has to be considered as a first step evaluation, waiting for better network connections, disk and systems when available at each site.

Responsability

SITE	ROLE/TASK
CINES	Full evaluation
CINECA, IDRIS, NIIF	Support to setup and configure iRODS servers for the workbench

References

- [1] https://www.irods.org/index.php/Release_Notes_3.2
- [2] https://www.irods.org/index.php/IRODS:Data_Grids,_Digital_Libraries,_Persistent_Archives,_and_Real-time_Data_Systems
- [3] <https://www.irods.org/index.php/glossary>
- [4] <https://www.irods.org/index.php/Downloads>
- [5] <https://bscw.zam.kfa-juelich.de/bscw/bscw.cgi/942629>
- [6] <http://fasterdata.es.net/host-tuning/linux>
- [7] <http://www.iozone.org/>

Tested Components

The transfer bandwidth performance evaluation focus on the iRODS software itself but also on the network [6] and I/O environment involved [7]. The related documentation on iRODS is specified at [1], [2], [3].

The iRODS release used for the evaluation is 3.2. It can be downloaded at [4].

Tested Features

The transfer bandwidth performance evaluation will focus on 3 major points:

- a) The test case : choice of the file and process
- b) The testbed : servers involved, iRODS servers, network configuration
- c) The tests

The methodology is the most important part of this work. This way, relevant results can be produced and compared.

Tests Environment

SITE	TEST ENVIRONMENT
CINES	SGI Altix XE 250 <ul style="list-style-type: none"> • 2 x Intel E5420 @ 2.5 GHz • 32 GB RAM

	<ul style="list-style-type: none"> • 1 Gb Ethernet card • 10 Gb Ethernet card • SUSE SLES 11 SP1 Local FS : Lustre @ ~577 MB/s R/W iRODS : 3.2
CINECA	DELL <ul style="list-style-type: none"> • 16 x Intel E5530 @ 2.4 GHz • 64 GB RAM • 1 Gb Ethernet card • Debian GNU/linux 6.0 Local FS : GPFS-NFS @ ~160 MB/s R/W iRODS : 3.2
IDRIS	IBM System x3655 <ul style="list-style-type: none"> • 4 x AMD Opteron 2218 @ 2.6 GHz • 2 GB RAM • 1Gb Ethernet card • Fedora release 17 (Beefy Miracle) Local FS : ext4 @ ~78 MB/s R/W iRODS : 3.2
NIIF	Cloud Virtual Machine <ul style="list-style-type: none"> • 2 cores • 2 GB • 100 Mb Ethernet card • Debian GNU/linux 6.0 Local FS : XFS @ ~13 MB/s R/W iRODS : 3.2

Methodology

The methodology used in this evaluation intends to use the methodology defined in the *New File Transfer Technology* task. However, this methodology was not fully applicable due to the reduced capacity testbed that was provided by some partners.

The initial methodology was defining the following set of information:

- Similar operational conditions (minimum requirements)
 - TCP buffer sizing
 - MTU and Jumbo Frame
 - Disk performance (no bottleneck)
 - Network Capacity (using PRACE dedicated 10 Gbe network)
- Specific Test Case
 - Dataset : A-small files, B- large files
 - Workload : A-100G, B-500G, C-1TB
 - Parallel Streams : A-1, B-4, C-8, D-16
- Performance reference value with gridFTP (in the same configuration)

Each test must be run several times to deliver a reliable measurement.

Test Case			
Run	Dataset type	workload	Parallel streams
1	A (1000 files of 100MB)	A (100GB)	A (4)
2	A (100 files of 1GB)	A (100GB)	A (1)

3	A (100 files of 1GB)	A (100GB)	A (4)
4	A (100 files of 1GB)	A (100GB)	B (8)
5	A (100 files of 1GB)	A (100GB)	C (16)
6	A (100 files of 5GB)	B (500GB)	A (4)
7	A (100 files of 5GB)	B (500GB)	B (8)
8	A (100 files of 5GB)	B (500GB)	C (16)
9	A (100 files of 10GB)	C (1000GB)	A (4)
10	A (100 files of 10GB)	C (1000GB)	B (8)
11	A (100 files of 10GB)	C (1000GB)	C (16)
12	B (1 file of 100GB)	A (100GB)	A (4)
13	B (1 file of 100GB)	A (100GB)	B (8)
14	B (1 file of 100GB)	A (100GB)	C (16)
15	B (1 file of 500GB)	B (500GB)	A (4)
16	B (1 file of 500GB)	B (500GB)	B (8)
17	B (1 file of 500GB)	B (500GB)	C (16)
18	B (1 file of 1TB)	C (1000GB)	A (4)
19	B (1 file of 1TB)	C (1000GB)	B (8)
20	B (1 file of 1TB)	C (1000GB)	C (16)

General Network Information				
Site	Capacity	RTT	Net.ipv4.tcp_rmem	Net.ipv4.tcp_wmem
CINECA	1 Gbps	14.1s	4194304 bytes	4194304 bytes
IDRIS	1 Gbps	27.1s	6291456 bytes	4194304 bytes
NIIF	1 Gbps	43.5s	4194304 bytes	4194304 bytes

This dataset case is able to provide a complete and detailed set of cases to analyze the performance of transfer tools, but with our test bed we were not able to use it because of the following three reasons:

- The main one is, only CINES provided an access to the dedicated high performance 10Gbe PRACE network. All other sites can only provide a public internet access to their iRODS server limited to a 1 Gbe shared link.
- Only CINES provided a gridFTP environment on the iRODS server to perform the reference measures.
- IDRIS and NIIF couldn't provide enough disk I/O performances on the test iRODS server they offered for the test.

Actually, the tests that were performed are the following ones:

Test Case			
Run	Dataset type	workload	Parallel streams
1	A (1 file of 100MB)	A (100MB)	A (1)
2	A (1 file of 100MB)	A (100MB)	A (4)
3	A (1 file of 100MB)	A (100MB)	A (8)
4	A (1 file of 100MB)	A (100MB)	A (16)
5	B (1 file of 1GB)	B (1GB)	B (1)
6	B (1 file of 1GB)	B (1GB)	B (4)
7	B (1 file of 1GB)	B (1GB)	B (8)
8	B (1 file of 1GB)	B (1GB)	B (16)

Tests Description

Once the iRODS servers 3.2 were deployed a “remote zone” was created on each iRODS servers to allow transfer of data. People involved in this work were granted access to the remote zone. These zones were created on the highest performance filesystem available on each server. (cf. *Tests Environment* above). The iRODS servers used the GSI certificate based authentication method for the users.

The two datasets used for the benchmark were created for the NFTT sub-task by a program using random number generator to build it contain to prevent from any compressing process during the transfer steps.

The iRODS servers were tuned regarding the network performances to ensure better performance in the transfer process. This has been done by adapting the following parameters in the server configuration file `~irods/server/config/reConfigs/core.re`

```
acSetNumThreads {msiSetNumThreads(sizePerThrInMb, maxNumThr, windowSize); }
```

- ***sizePerThrInMb*** : The number of threads is computed using: $numThreads = fileSizeInMb / sizePerThrInMb + 1$ where *sizePerThrInMb* is an integer value in MBytes. It also accepts the word “default” which sets *sizePerThrInMb* to a default value of 32
- ***maxNumThr*** : The maximum number of threads to use. It accepts integer value up to 16. It also accepts the word “default” which sets *maxNumThr* to a default value of 4. A value of 0 means no parallel I/O. This can be helpful to get around firewall issues.
- ***windowSize*** : the tcp window size in Bytes for the parallel transfer. A value of 0 or “default” means a default size of 1,048,576 Bytes.

The runs were done using a dedicated script program which performed network performance measurement before each run and executed the transfer using the iRODS *iput* command several time to ensure reliability on the performance printed out.

Final results

Run	Dataset type	workload	Parallel streams	CINECA	IDRIS	NIIF
1	A (1 file of 100MB)	A (100MB)	A (1)	0.26 MB/s	0.63 MB/s	0.07 MB/s
2	A (1 file of 100MB)	A (100MB)	A (4)	1.02 MB/s	1.08 MB/s	0.28 MB/s
3	A (1 file of 100MB)	A (100MB)	A (8)	2.22 MB/s	1.09 MB/s	0.53 MB/s
4	A (1 file of 100MB)	A (100MB)	A (16)	3.03 MB/s	1.09 MB/s	0.85 MB/s
	B (1 file of 1GB)	B (1GB)	Server choice	6.89 MB/s	2.24 MB/s	1.88 MB/s
5	B (1 file of 1GB)	B (1GB)	B (1)	1.07 MB/s	1.30 MB/s	0.147 MB/s
6	B (1 file of 1GB)	B (1GB)	B (4)	2.06 MB/s	2.24 MB/s	0.579 MB/s
7	B (1 file of 1GB)	B (1GB)	B (8)	5.27 MB/s	2.24 MB/s	1.08 MB/s
8	B (1 file of 1GB)	B (1GB)	B (16)	6.02 MB/s	2.24 MB/s	1.90 MB/s

These numbers show that:

- CINECA transfer with iRODS is able to provide good performance up to 70% of the peak of its opened and shared internet network link.
- IDRIS iRODS server is limited by the I/O bottleneck on its server.
- NIIF iRODS server is limited by the bad network performances and the poor disk I/O rate and the server resources (virtual machine with reduced resources: cpu, memory, network, and disk I/O)
- Threads specification at runtime as a parameter to the *iput* command do not give better results than default iRODS settings because of the good tuning of the iRODS server configuration.

Conclusions

The tuning is always an important part of the work when talking about performance. Once the network was correctly tuned the iRODS tuning part was simple and easy to do and iRODS was then able to provide good performance without any runtime setting.

iRODS appears here as a simple tool for transferring files allowing good performance. Performance tests should be continued including additional test cases on the PRACE high performance 10 Gb/s dedicated network to show the full iRODS performance capacity.

Despite the testbed characteristics which were not fitting with the performance goals of this work, iRODS has shown that it was able to provide good performances. It stays a real and good challenger to the standard transfer tool gridFTP offering by the way much more powerful functionalities than only transfer with a simple filesystem like command approach.

6.6.6 iRODS - Direct Access Resources Evaluation

Introduction

This document is within the scope of the iRODS sub-task in T10.2 “Evaluating data services”. It provides a homogeneous way to evaluate features and software for a better coherency within the iRODS working group. This document evaluates the Direct Access Resources feature of iRODS 3.2.

Purpose

The iRODS organizes its storage locations as resources. The Direct Access resources feature provides support for sharing access to a storage location, i.e. resource, with the regular filesystem access. A typical usage scenario would be an environment in which there is a shared high performance file system mounted on a compute cluster via NFS, and on which iRODS has the files from this file system registered in order to provide meta-data annotation for the files in this file system (i.e. iRODS acts as an "overlay" for the UNIX file system).

To make this possible, the system relies on identical user ids and passwords on the iRODS service and user clients. Also this feature relies on the filesystem metadata feature being enabled during iRODS installation.

This evaluation will test the basic functionality of creating the direct access resources and accessing the files, comparing what is seen through iRODS access with direct filesystem access.

Responsibility

SITE	ROLE/TASK
IPB	Full evaluation

References

List the references, applicable documents and related documentation (user, technical, ...)

- [1] https://www.irods.org/index.php/Release_Notes_3.2
- [2] https://www.irods.org/index.php/Direct_Access_Resources
- [3] https://www.irods.org/index.php/File_System_Meta-data
- [4] https://www.irods.org/index.php/Run_server_as_root
- [5] <https://groups.google.com/forum/#!topic/iROD-Chat/Q3MvceznE3E>
- [6] <https://www.irods.org/index.php/glossary>
- [7] <https://www.irods.org/index.php/Downloads>

Tested Components

This document evaluates the Direct Access Resources feature of iRODS 3.2.

Tested Features

This evaluation focuses on the file access through iRODS and directly through the file system, which is provided by the Direct Access Resource feature [2].

Non Tested Features

All other iRODS features that are not directly relevant for the evaluation of the Direct Access Resources.

Tests Phases

There were two types of tests performed:

- checking if the direct access works for resources on the local filesystem
- checking if the access also works for shared filesystem mounted through NFS.

For each of the test cases, the following subtests were executed:

- testing if the resource creation is working
- testing if the file creation and the access through the filesystem and the resource are working.

Tests Environment

The iRODS server and clients (icommands) had to be build with the following build flags enabled in config/config.mk:

- FILESYSTEM_META = 1
- RUN_SERVER_AS_ROOT = 1
- DIRECT_ACCESS_VAULT = 1

as described in [2], [3] and [4].

The server was started as the root user, while database (default PostgreSQL, that comes bundled with the installation) had to be run as a regular user. IRODS users had to have accounts on the host machine with the same username and password in order to have access to files through the filesystem.

SITE	TEST ENVIRONMENT
IPB	Scientific Linux 6.3 virtual machine on PARADOX 2 Proc Intel Xeon CPU E5345, 2.33GHz 2 GB RAM 1Gb Ethernet card

Testing Methodology

The test will focus on the functional aspect. So, we will check whether the functionality is provided that if whether it does what is supposed to do, the easiness of use, the reliability and maintainability.

Tests Description

Direct access resource creation

Initial conditions	<ul style="list-style-type: none"> • irods user account in iRODS, defined as "rodsadmin". • irods service running as root, with DIRECT_ACCESS_VAULT configuration enabled. • MySQL used for ICAT database, started independent of iRODS, or an independent PostgreSQL instance must be used because default PostgreSQL that is bundled with iRODS can not be started as root <ul style="list-style-type: none"> ◦ alternatively when starting the server as root use istart command instead of start, passed to irodsetl script.
Parameters and input data	<ul style="list-style-type: none"> • Direct access resource at path: /opt/rodshare • a dummy text file used to verify that resource is accessible for file operations.

Test procedure	<ul style="list-style-type: none"> • cd \$IRODS_HOME/ • sudo ./irodsctl start • iinit • iadmin mkresc testresc "direct access filesystem" cache irods.ipb.ac.rs /opt/rodshare
Expected result	testresc resource should be created on the irods instance.
Test result	<pre>\$ iadmin lr demoResc testresc \$</pre> <p>The test result matches the expected result.</p>

Direct access resource file creation and access through iRODS and through local filesystem

Initial conditions	<ul style="list-style-type: none"> • "testresc" direct access resource created (see 11.1 Direct access resource creation) • user pr1ig000 should be created on the unix domain in which irods instance is created and in the irods service itself
Parameters and input data	<ul style="list-style-type: none"> • Direct access resource at path: /opt/rodshare • a dummy text file used to verify that resource is accessible for file operations.
Test procedure	<ul style="list-style-type: none"> • iinit (enter pr1ig000's password) • iput -R testresc testfile.txt
Expected result	The user which created the file and put it into irods should also be the owner of the file in the local filesystem on the resource location.
Test result	<pre>\$ ils /IPB/home/pr1ig000: testfile.txt \$ cd /opt/rodshare/home/pr1ig000 \$ ls -l File: `testfile.txt' Size: 124 Blocks: 8 IO Block: 4096 regular file Device: 803h/2051d Inode: 2501436 Links: 1 Access: (0664/-rw-rw-r--) Uid: (501/pr1ig000) Gid: (501/pr1ig000) Access: 2013-07-03 19:13:34.131999965 +0200 Modify: 2013-07-03 19:13:34.131999965 +0200 Change: 2013-07-03 19:13:34.131999965 +0200 \$ cat testfile.txt file contents... \$</pre> <p>The test result matches the expected result. (Uid and Gid of the file match the user who put the file</p>

	into the resource)
--	--------------------

Adding metadata to file in direct access resource

Initial conditions	<ul style="list-style-type: none"> See 11.2 <i>Direct access resource file creation through iRODS and local filesystem</i>
Parameters and input data	<ul style="list-style-type: none"> Direct access resource at path: /opt/rodshare a dummy text file (testfile.txt) used to verify that resource is accessible for file operations. Metadata triplet added has following details <ul style="list-style-type: none"> AttrName: "TextType" AttrValue: "random" AttrUnit: <i>none</i>
Test procedure	<ul style="list-style-type: none"> iinit (type in pr1ig000's password) imeta add -d testfile.txt TextType random
Expected result	The test file should have metadata record associated with it.
Test result	<pre>\$ imeta ls -d testfile.txt AVUs defined for dataObj testfile.txt: attribute: TextType value: random units: \$</pre> <p>The test result matches the expected result.</p>

Direct access resource creation on NFS

Initial conditions	<ul style="list-style-type: none"> See 11.2 <i>Direct access resource file creation through iRODS and local filesystem</i> nfs mounted on /nfs
Parameters and input data	<ul style="list-style-type: none"> Direct access resource at path: /nfs The resource created is named "nfstestresc"
Test procedure	<ul style="list-style-type: none"> iinit (enter rodsadmin's password) iadmin mkresc ntfresc "direct access filesystem" cache irods.ipb.ac.rs /nfs
Expected result	nfstestresc resource should be created on the irods instance.
Test result	<pre>\$ ilsresc demoResc testresc nfstestresc \$</pre> <p>The test result matches the expected result.</p>

File access and creation through iRODS and filesystem in direct access resource on NFS

Initial conditions	<ul style="list-style-type: none"> See 11.2 <i>Direct access resource file creation through iRODS and local filesystem</i> nfs mounted on /nfs
---------------------------	--

Parameters and input data	<ul style="list-style-type: none"> • Direct access resource at path: /nfs • The resource created is named "nfstestresc" • a dummy text file used to verify that resource is accessible for file operations.
Test procedure	<ul style="list-style-type: none"> • iinit (enter pr1ig000's password) • iput -R nfstestresc testfile.txt
Expected result	testfile.txt should be created on /nfs/home/pr1ig000 path, and it should have pr1ig000 as the file owner.
Test result	<pre>\$ ls -l testfile.txt pr1ig000 0 nfstestresc 0 2013-07-04.11:13 & testfile.txt \$ cd /nfs/home/pr1ig000 \$ ls -l testfile.txt -rw-r--r--. 1 root root 124 Jul 4 11:13 testfile.txt \$</pre> <p>The test result does not match the expected result.</p> <p>The documentation [2] explains that this is because the irods user which put the file does not have write permissions on the filesystem location of the resource. But, it does not work even after giving permissions to the user and making him an owner of the /nfs/home/pr1ig000 subdirectory.</p>

Conclusions

The Direct Access Resource feature provides a way to have direct access to the files in a resource through the filesystem they reside on. However, the feature depends on iRODS users having the same accounts on the machine that hosts the filesystem, and having sufficient file access rights. If the access rights are lacking for a given user, the files on the system will be owned by the root user.

Since the iRODS server must run as root for the direct access resources to work, the need for this feature should be carefully weighed against possible security concerns.

6.6.7 iRODS iDROP evaluation

Introduction

This document is within the scope of the iRODS sub-task in T10.2 "Evaluating data services". It provides a homogeneous way to evaluate features and softwares for a better coherency within the iRODS working group. This document evaluates the iDrop Graphical User Interface and its functionality within the iRODS data grid.

Purpose

iDrop is a user-friendly desktop GUI that manages data movement and synchronization. It provides a graphical view of an iRODS data grid, supporting drag and drop transfers between iRODS and the local file system, as well as data movement within an iRODS grid. iDrop uses the Jargon-core client libraries to establish a direct connection to iRODS via the iRODS XML client protocol. This provides for efficient data transfer, including the ability to use the parallel data transfer algorithm.

Using the iDrop GUI, users can:

- Put files to iRODS from the local file system.
- Get files from iRODS to the local file system.
- Create/delete/rename iRODS files.
- Copy and move files in iRODS.
- Replicate iRODS files.
- Manage the automated synchronization of directories between the local file system and the iRODS data grid.

The iDrop-web interface is a suite of tools that provide individuals and groups functionality of iDrop through a web service. The idrop-web interface includes the idrop-lite Java applet for bulk uploads and downloads, and provides Java Web Start links to launch the iDrop desktop GUI. It is deployed as a standard .war file on any commodity Java container that supports the servlet specification, such as Apache Tomcat.

iDrop is mainly end-user orientated software, although some of its functionality could be useful to administrators too.

Responsability

SITE	ROLE/TASK
IPB	Full evaluation

References

- [1] <https://code.renci.org/gf/project/irodsidrop/>
 [2] <http://www.java.com/>
 [3] <http://tomcat.apache.org/>

Tested Components

iDrop 2.0.0 Release

iDrop contains web and client GUI for interacting with iRODS:

- iDrop Swing GUI - transfer and synchronization manager
- iDrop Web Interface personal cloud web interface
- iDrop Lite applet - transfer applet for embedding in iDrop Web

Significant components:

- idrop.jnlp (<http://iren-web.renci.org/idrop-release/idrop.jnlp>) – iDrop Web Start Application for iDrop desktop GUI
- idrop-web2.war (<https://code.renci.org/gf/download/frsrelease/157/1229/idrop-web2.war>) – pre-compiled war file for deploying iDrop Web Interface
- idrop-lite-2.0.0-jar-with-dependencies.jar (<https://code.renci.org/gf/download/frsrelease/157/1228/idrop-lite-2.0.0-jar-with-dependencies.jar>) - transfer applet for embedding in iDrop Web Interface

Tested Features

iDrop desktop and web interfaces.

Non Tested Features

Other features of iRODS are not part of this evaluation.

Tests Phases

- Phase 1 included setting up the testing environment and enablement of iDrop Web Interface service
- Phase 2 included exploration and testing of iDrop features and functionality

Tests Environment

SITE	
IPB	
iRODS server	KVM Virtual Machine on PARADOX Scientific Linux 6.3 2 Proc Intel Xeon CPU E5345, 2.33GHz 2 GB RAM 1Gb Ethernet card
Client	Microsoft Windows 7 Home Premium Ubuntu 12.04 Java SE Update 25 (on both OSes) Apache Tomcat 6.0 (on both OSes)

iDrop explicitly requires username and password for authentication, so it is important to note that iDrop doesn't work with GSI authentication. Therefore, this testing was performed on an iRods 3.2 instance without GSI support.

Testing Methodology

Evaluation was performed as a series of feature tests following the test script adopted and revised from the iDrop project home site:

<https://code.renci.org/gf/project/irodsidrop/wiki/?pagename=iDrop+testing+script>.

The goal of this evaluation is to get familiar with iDrops features and to examine its functionality, ease of use, reliability and efficiency.

Tests Description

For each group of functional features, descriptions and comments will be given if needed.

iDrops Desktop (Swing) GUI

Login <ul style="list-style-type: none"> • Cancel login • Bad host/uid/password • Launching second instance 	Login screen is simple and functional, with all the necessary fields and responds to all stated situations in a proper manner and with proper notifications
Local Tree <ul style="list-style-type: none"> • Browsing and selecting local files and folders and getting relevant info • Creating/renaming/deleting local folders • Recursive deleting 	Standard desktop functionality. Path, size and last modification time displayed when cursor is put over a file or folder in the local tree.
iRods Tree	Same functionality as with the Local Tree but with notable lag. Uploads and downloads can

<ul style="list-style-type: none"> • Browsing and selecting files and folders and getting relevant info • Creating/renaming/deleting local folders • Upload and download of files and folders • Progress bar • Refresh button • Tree root 	<p>be performed via drag&drop or interface buttons. Progress bar indicates transfers. When uploading and downloading empty folders, status bar doesn't indicate progress but remains at 0%. Refresh button exists but it is usually not necessary as iRODS tree refreshes by itself. Refresh function maintains expansion of the file tree. There is a drop down menu for setting the tree root for easier navigation.</p>
<p>Copy & Move</p> <ul style="list-style-type: none"> • Option key for drag&drop 	<p>Ctrl key can be used as an option key for drag&drop. When pressed, copy function is executed and move function otherwise. There is also an interface button for copy/move.</p>
<p>Info panel</p> <ul style="list-style-type: none"> • Tags and comments • Metadata • Permissions 	<p>Interface button brings up an info panel for the current selection. Tags and comments can be updated. Metadata can be created and/or deleted. User can set permissions for the current selection within his rights.</p>
<p>Search</p>	<p>Only by filename.</p>
<p>Desktop – iRODS drag&drop</p> <ul style="list-style-type: none"> • Desktop to iRODS • iRODS to Desktop • Option key 	<p>Recursive directory and files drag&drop from desktop to iRODS and from iRODS to desktop works with the same option key functionality.</p>
<p>Settings</p> <ul style="list-style-type: none"> • iDrop • Accounts • Transfers • Synchronization 	<p>Four tabs. Show iDrop GUI on startup and Show within-file transfer progress can be checked in iDrop tab. Accounts tab gives options for Default Resource, Login to Another Grid and Change Password. Transfers tab gives options for Transfer Management, Parallel Transfer Options, Buffer Options and settings for iRODS agent connection timeout, with options to restore default settings.</p>
<p>Synchronization</p> <ul style="list-style-type: none"> • Synchronization mode • Synchronization frequency • Status bar 	<p>List of set synchronizations is displayed with appropriate folder paths. Only local to iRODS synchronization mode is operational. There is placeholder for iRODS to local and bidirectional mode but they are still not implemented. Drop down menu for setting Synchronization frequency has only four values: Hourly, Weekly, Daily and Every two minutes for testing purposes. Set synchronizations from the list can be forced to synchronize. Status bar on the main window of the interface also indicates status of synchronization.</p>

System tray icon	iDrop Desktop edition places an icon in the system tray with the common set of options when right-clicked.
-------------------------	--

iDrops Web Interface

Login and Home screen <ul style="list-style-type: none"> • Starred Files • Starred Folders • Folders shared by me • Folders shared with me • Quick upload 	Login screen can be modified through idrop-web.config2.groovy file in /etc/idrop-web directory. Default values for host, port, zone, resource and authorization scheme can be set and they won't be displayed on the login screen. Home screen offers overview of starred files and folders, shared folders and quick upload tool that uploads selection to predefined folder.
Browse screen <ul style="list-style-type: none"> • Tree context menu • Add to cart • Bulk upload • Tickets 	Browse screen offers iRODS tree view with very user friendly interface. Tree context menu contains all the options for refreshing, creating, renaming, deleting, cut/copy/pasting of the content and getting corresponding information. Uploading can be done by Quick upload or Bulk upload for multiple selections. Download is managed by the shopping cart feature that lets you store your choices and download them at any time by checking out. Info view gives basic information as well as information on tags and metadata with editing and updating options. There is also a Ticket feature tab. Tickets are tokens to iRODS files and collections that may be shared. Anyone with a ticket may access your data, so you can email them or share them on social media sites. There is also an option to mark files or folders as starred and these can be viewed on the appropriate link on the Home screen.
Profile	This screen provides options for entering additional information about the user.
Search	Search files and folders by tags. Search results can be deleted or added to the Shopping cart for download.
Tools	Tools option provides link to iDrop desktop application.
Account	Logout, Change password and Set default resource options
Shopping Cart	Beside already mentioned functionality, shopping cart offers options for clearing, deleting and reloading added items.

Discussion and Conclusions

This document focused on iDrop features and user experience and not on the setting of the testing environment or its integration with iRODS and possible technical issues.

The iDrop desktop GUI is a useful tool but it still has much place for improvement. The main problem is lag which doesn't happen when you browse local files and that is not an essential feature from the iRODS perspective. Because of that usage can be quite difficult. There is also an issue of limited search and authentication options. On the other hand, the iDrop Web Interface is fast, intuitive and easy to use. It also shares limited search options and authentication problem but it broadens its options by including a direct link to the iDrop Desktop GUI to complement some of its flaws. To get the most of the iDrop functionality, Desktop and Web interface should be used together. Hopefully, further development and future versions will make this GUI for iRODS an obvious choice for both end-users and administrators but for the time being, mostly due to lack of support for GSI, its use is limited.