SEVENTH FRAMEWORK PROGRAMME
Research Infrastructures

INFRA-2010-2.3.1 – First Implementation Phase of the European High Performance Computing (HPC) service PRACE

PRACE-1IP

PRACE First Implementation Project

Grant Agreement Number: RI-261557

D6.2
First annual report on the technical operation and evolution

Final

Version:	1.0
Author(s):	Gabriele Carteni (BSC), Giuseppe Fiameni (CINECA)
Date:	27.06.2011

Project and Deliverable Information Sheet

PRACE Project	**Project Ref. №: RI-261557**	
	Project Title: PRACE First Implementation Project	
	Project Web Site: http://www.prace-project.eu	
	Deliverable ID: D6.2	
	Deliverable Nature: DOC_TYPE: Report	
	Deliverable Level: PU	**Contractual Date of Delivery:** 30 / June / 2011
		Actual Date of Delivery: 30 / June / 2011
	EC Project Officer: Bernhard Fabianek	

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

Document Control Sheet

Document	**Title: First annual report on the technical operation and evolution**	
	ID: D6.2	
	Version: 1.0	**Status:** Final
	Available at: http://www.prace-project.eu	
	Software Tool: Microsoft Word 2008 for Mac (v.12.0.0)	
	File(s): D6.2.doc	
Authorship	**Written by:**	Gabriele Carteni (BSC), Giuseppe Fiameni (CINECA)
	Contributors:	Frank Scheiner (HLRS), Stephanie Meier (FZJ), Ralph Niederberger (FZJ), Michael Rambadt (FZJ), Jutta Docter (FZJ), Jules Wolfrat (SARA), Morgotti Marcello (CINECA), Xavier Delaruelle (CEA), Miroslaw Kupczyk (PSNC), Liz Sim (EPCC), Ilya Saverchenko (LRZ), Miroslaw Kupczyk (PSNC), Denis Girou (IDRIS), Ioannis Liabotis (GRNET), Apollon Oikonomopoulos (GRNET), Jarno Laitinen (CSC), Axel Berg (SARA)
	Reviewed by:	Dietmar Erwin (FZJ) Jussi Heikonen (CSC)
	Approved by:	Technical Board

Document Status Sheet

Version	Date	Status	Comments
0.1	15/05/2011	Draft	Outline
0.2	26/05/2011	Draft	First internal release. Added first contributions

			from subtask leaders.
0.2.1	26/05/2011	Draft	Added information about CURIE System to Ch.2
0.3	01/06/2011	Draft	- Added sections 3.1, 3.6, 3.7 - Added sections 4.5, 4.7, 4.8 - Updated sections 3.3, 4.1 - Draft released for internal WP6 review
0.4	09/06/2011	Draft	- Updated contributors list - Merged comments of contributors (CINECA, LRZ) - Added Introduction and Conclusion - Added chapter 5
0.5	10/06/2011	Draft (ready for internal review)	- Review on Chapter 4 - New contribution for technical requirements definition - Conclusion
0.6	28/06/2011	Draft	- Merged PRACE reviewers' comments
1.0	27/06/2011	Final Version	

Document Keywords

Keywords:	PRACE, HPC, Research Infrastructure, Operations, Software Catalogue, Deployment, Common Services, Tier-0, Tier-1, User Support, Technology Assessment, Service Category, Resources Integration

Copyright notices

Table of Contents

List of Figures

List of Tables

References and Applicable Documents

[1] PRACE Project: http://www.prace-project.eu
[2] PRACE Research Infrastructure AISBL: http://www.prace-ri.eu
[3] PRACE Wiki - WP6: https://prace-wiki.fz-juelich.de/bin/view/PRACE/WP6/WebHome
[4] IGTF: http://www.igtf.net/
[5] EUGridPMA: http://www.eugridpma.org
[6] Internet X.509 Public Key Infrastructure Certificate and Certificate Revocation List (CRL) Profile RFC 5280: https://datatracker.ietf.org/doc/rfc5280
[7] Distribution of CA information: http://winnetou.sara.nl/deisa/certs
[8] Fetch-crl utility: https://dist.eugridpma.info/distribution/util/fetch-crl3
[9] OpenSSH: http://www.openssh.org
[10] Globus GSI-OpenSSH: http://www.globus.org/toolkit/security/gsiopenssh
[11] Usage Record – Format recommendation: http://www.ogf.org/documents/GFD.98.pdf
[12] Accounting Facilities in the European Supercomputing Grid DEISA, J. Reetz, T. Soddemann, B. Heupers, J. Wolfrat eScience Conference 2007 (GES 2007) http://www.ges2007.de/fileadmin/papers/jreetz/GES_paper105.pdf
[13] DART: http://www.deisa.eu/usersupport/user-documentation/deisa-accounting-report-tool
[14] Inca: http://inca.sdsc.edu/drupal/
[15] Grid-SAFE: http://gridsafe.forge.nesc.ac.uk/Documentation/GridSafeDocumentation/
[16] EU directive 95/46/EC: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31995L0046:EN:HTML
[17] EMI: http://www.eu-emi.eu
[18] IGE: http://www.ige-project.eu
[19] SCHAC schema: http://www.terena.org/activities/tf-emc2/schacreleases.html
[20] UNICORE: http://www.unicore.eu/
[21] TeraGrid: http://www.teragrid.org

[22] EGI (European Grid Infrastructure): http://www.egi.eu
[23] Non-functional requirements, http://en.wikipedia.org/wiki/Non-functional_requirement
[24] M. Glinz. Rethinking the Notion of Non-Functional Requirements, http://www.ifi.uzh.ch/rerg/fileadmin/downloads/publications/papers/3WCSQ2005.pdf
[25] Martin Glinz, "On Non-Functional Requirements," re, pp.21-26, 15th IEEE International Requirements Engineering Conference (RE 2007), 2007
[26] *ISO/IEC 9126:* http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=22749
[27] GEANT Project: http://www.geant.net
[28] Globus Online: https://www.globusonline.org
[29] P. Kunszt, "Requirements Analysis for Tier-0 Systems Management", PRACE-PP D4.1.1
[30] R. Murri, "Deployment of enhanced solutions", PRACE-PP D4.2.2
[31] Iperf: http://iperf.sf.net/
[32] ProActive Parallel Suite: http://proactive.inria.fr
[33] OW2 Consortium: http://www.ow2.org
[34] Coordinated TeraGrid Software and Services (CTSS): http://www.teragrid.org/web/user-support/ctss
[35] Distributed European Infrastructure for Supercomputing Applications (DEISA): http://www.deisa.eu
[36] TeraGrid Common User Environment (CUE): https://www.teragrid.org/web/user-support/cue-getting-started
[37] Modules: http://modules.sourceforge.net
[38] SoftEnv: http://www.mcs.anl.gov/hs/software/systems/softenv/softenv-intro.html
[39] DEISA User Documentation: http://www.deisa.eu/usersupport/user-documentation
[40] LRZ Remote Visualization Guide: http://www.grid.lrz.de/de/mware/globus/client/gsissh_term_visualisation.html
[41] LRZ Applications available on the remote visualisation servers: http://www.lrz.de/services/compute/visualisation/visualisation_5/index.html
[42] Miroslaw Kupczyk, "Specification for PRACE systems management", PRACE-PP D4.3
[43] GridFTP: http://globus.org/toolkit/docs/3.2/gridftp
[44] iRODS: http://www.irods.org
[45] JuBE, Jülich Benchmarking Environment: http://www2.fz-juelich.de/jsc/jube
[46] PRACE SVN: https://subtrac.sara.nl/prace/svn/
[47] PRACE 1IP deliverable D6.1 Assessment of PRACE operational structure, procedures and policies
[48] PRACE 1IP deliverable D4.1 PRACE User Forum
[49] dCache: http://www.dcache.org
[50] OpenStack: http://www.openstack.org
[51] Storage Resource Manager (SRM) Interface specification: https://sdm.lbl.gov/srm-wg/doc/SRM.v2.2.html

List of Acronyms and Abbreviations

AAA	Authentication, Authorization and Accounting
AUP	Acceptable Use Policy
BSC	Barcelona Supercomputing Center (Spain)
BSS	Batch Scheduler System
BWCTL	Bandwidth Test Controller
CA	Certificate Authority

CEA	Commissariat à l'Energie Atomique et aux Energies Alternatives (contributor of GENCI, France)
CGI	Common Gateway Interface
CINECA	Consorzio Interuniversitario, the largest Italian computing centre (Italy).
CINES	Centre Informatique National de l'Enseignement Supérieur (contributor of GENCI, France)
CP/CPS	Certification Policy and Certification Practice Statement
CRL	Certificate Revocation List
CSC	Finnish IT Centre for Science (Finland)
DART	DEISA Accounting Report Tool
DEISA	Distributed European Infrastructure for Supercomputing Applications. EU project by leading national HPC centres.
EGI	European Grid Infrastructure
EMI	European Middleware Initiative
EPCC	Edinburgh Parallel Computing Centre (represented in PRACE by EPSRC, United Kingdom)
FZJ	Forschungszentrum Jülich (Germany)
GB	Giga (= 2^{30} ~ 10^{9}) Bytes (= 8 bits), also GByte
GB/s	Giga (= 10^{9}) Bytes (= 8 bits) per second, also GByte/s
GFlop/s	Giga (= 10^{9}) Floating point operations (usually in 64-bit, i.e. DP) per second, also GF/s
GCS	Gauss Centre for Supercomputing (Germany)
GENCI	Grand Equipement National de Calcul Intensif (french representative in PRACE, France)
GHz	Giga (= 10^{9}) Hertz, frequency =10^{9} periods or clock cycles per second
GSI-SSH	A ssh client and server implementation using X.509 certificates with OpenSSH
HLRS	High Performance Computing Center Stuttgart (Germany)
HPC	High Performance Computing; Computing at a high performance level at any given time; often used synonym with Supercomputing
HSM	Hierarchical Storage Management
HTC	High Throughput Computing
IDRIS	Institut du Développement et des Ressources en Informatique Scientifique (contributor of GENCI, France)
IGE	Initiative for Globus in Europe
IGTF	International Grid Trust Federation
IPB	Institute of Physics, Belgrade (Serbia)
ISTP	Internal Specific Targeted Projects
KTH	Kungliga Tekniska Högskolan (represented in PRACE by SNIC, Sweden)
LDAP	Lightweight Directory Access Protocol
LRZ	Leibniz Supercomputing Centre (Garching, Germany)
MB	Mega (= 2^{20} ~ 10^{6}) Bytes (= 8 bits), also MByte
MB/s	Mega (= 10^{6}) Bytes (= 8 bits) per second, also MByte/s
NTNU	Norwegian University of Science and Technology (Trondheim, Norway)
OGF	Open Grid Forum
PCPE	PRACE Common Production Environment
PFlop/s	Peta (= 10^{15}) Floating-point operations (usually in 64-bit, i.e. DP) per second, also PF/s
PI	Principal Investigator - the coordinator for project proposals
PKI	Public Key Infrastructure

PMA	Policy Management Authority
PRACE	Partnership for Advanced Computing in Europe; Project Acronym
PRACE-PP	PRACE Preparatory Phase Project
PRACE-RI	PRACE Research Infrastructure
PRACE-1IP	PRACE First Implementation Phase
PRACE-2IP	PRACE Second Implementation Phase
PSNC	Poznan Supercomputing and Networking Centre (Poland)
RI	Research Infrastructure
SAML	Security Assertion Markup Language
SNIC	Swedish National Infrastructure for Computing (Sweden)
SRM	Storage Resource Management
STS	Security Token ServiceX.509 An US infrastructure combining leadership class resources at 11 partner sites to create an integrated, persistent computational resource
TFlop/s	Tera (= 10^{12}) Floating-point operations (usually in 64-bit, i.e. DP) per second, also TF/s
TGCC	Très Grand Centre de calcul du CEA. CEA Very Large Computing Centre installed on the site of Bruyères-le-Châtel, France
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
TLS	Transport Layer Security
TSI	UNICORE Target System Interface
UiB	University of Bergen (Norway)
UNICORE	Uniform Interface to Computing Resources
UR-WG	Usage Record Working Group of the OGF
X.509	A format for the storage of identity information together with a public key as used in public key or asymmetric encryption

Executive Summary

Since the start of the PRACE-1IP project in July 2010 the Tier-0 service at FZJ in Germany has been available. Since February 2011 the second Tier-0 system at CEA in France has become available for users. We have setup and deployed a set of common operational services that integrate these Tier-0 services and that are prepared to integrate other Tier-0 services in the near future. Through this set of common services PRACE is presented to its users as a single coordinated distributed research infrastructure and allows users to use the PRACE distributed research infrastructure as seamlessly as possible.

Services that have been deployed include network services (e.g., Iperf), compute services (e.g., UNICORE), data services (e.g., GridFTP), AAA services (e.g., central LDAP, PRACE Accounting services, GSI-SSH), monitoring services (e.g., Inca), and user services (e.g., PRACE Common Production Environment, PRACE Help Desk). This set of services is primarily based on the work done in the PRACE Preparatory Phase project and on experiences from the DEISA2 project [35].

For the evolution of services and the deployment of new services user requirements are the most determining factor. We have performed an extensive user requirement analysis by means of a user survey. From this survey we have extracted high level requirements and we have defined corresponding plans to address those requirements. Besides we have performed technology watch in various service and technology domains, thereby also collaborating and relying in this first year on the DEISA2 technology assessment activities. Together this has led to a plan for technology assessment in year two of the project. Processes have been defined for service certification.

1 Introduction

Building and maintaining of the PRACE distributed Research Infrastructure is a continuous process that requires much interaction between distinct domains: technical, legal, users, vendors and operating staff. These entities interact with each other continuously and are essential for the definition of a properly working computing environment.

This document is the first annual report on activities carried out and outcomes achieved within WP6 – "Technical Operation and Evolution of the Distributed Infrastructure" of the PRACE-1IP project.

WP6 focuses on the PRACE Operational service provision for the users to allow them use the distributed research infrastructure as seamlessly as possible, and to deploy and develop services that are sustainable, of high quality, up-to-date and which fulfil the needs and requirements of the different users and user communities.

The common mission is to present PRACE to the users as a single distributed research infrastructure, instead of a set of individual systems/computing centres. To achieve this goal, the operational work has been divided and organised along two main tracks:

- Selection and deployment of common services, to assure and maintain a fully operational infrastructure (Task 6.2);
- Evaluation of new technologies, to improve performance and capabilities with cutting-edge advances in High Performance Computing and to fulfil new requirements coming from different scientific communities (Task 6.3).

The structure of this report reflects the organization of the work package and the outcomes produced by its subtasks. Policies and procedures, which have been adopted for coordinating all the operational activities, are described in the deliverable D6.1[47], published in conjunction with this document.

After this first year, two Tier-0 systems are fully available for production runs in PRACE:

- JUGENE, a IBM Blue Gene/P based machine installed at FZJ (GCS);
- CURIE, a BULL bullx system-based cluster installed at CEA (GENCI)

Actions regarding HERMIT, the CRAY XE6 system to be the third Tier-0 production system in PRACE and installed at HLRS (GCS), are not included in this report.

All selected software have been organised in a software catalogue, which acts as the main reference for the selection and deployment process. They have been selected by starting from evaluations made in the preparatory phase of PRACE and the large and consolidated experience handed over by seven years of the DEISA project phases. In both cases the different nature of the underlying infrastructure has been also considered (HPC infrastructure versus system prototypes in PRACE-PP versus the Grid infrastructure in DEISA).

Integration of the current Tier-0 systems, and the preparation for further Tier-0 integrations, has been realized by supporting directly Tier-0 centres with documentation for installation and configuration of software belonging to the following six service categories:

- Network Services
- Data Services
- Compute Services

- AAA (Authentication Authorization Accounting) Services
- User Services
- Monitoring Services

Besides the integration of existing Tier-0 systems under a common operation domain, WP6 has been working to assess and enhance the technological layer of the PRACE infrastructure towards a well-integrated, user-driven set of useful services. On the basis of collected user requirements an evolution plan has been defined for each service category in order to ensure that measureable results are delivered to users in a reasonable amount of time.

1.1 Background and purpose

The first year of the first implementation phase of PRACE (PRACE-1IP) started by focusing on requirements to integrate Tier-0 systems, which are high-end computing systems, in order to create a single infrastructure. In addition the future integration of Tier-1 systems has been considered, this integration will start on September 2011 in conjunction with the beginning of the second implementation phase of PRACE (PRACE-2IP).

An important background strategy was to adopt a top-down approach by fulfilling first all requirements for the integration of Tier-0 systems and then focusing on integration with Tier-1 systems.

However, in this first year, requirements for Tier-1 systems have been anticipated by checking the outcomes of DEISA, which is a valuable reference for the future Tier-1 infrastructure. All service implementations reflect this evaluation in order to assure a smooth integration between Tier-0 and Tier-1 sites.

Solutions have been proposed to allow seamless access to computing services of PRACE Tier-0 systems, to provide reliable data exchange mechanisms, to assure quality control for services provisioning as well as to offer users with a homogeneous interaction layer. The initial software stack covers the areas of User Administration and Accounting, Resource Management and Access, Distributed Data Management, and Monitoring of Distributed Resources.

Solutions for supporting end-users (first level) as well as system administrators (second level) have been defined with respect to the user support model described in the deliverable D6.1.

Accounting strategies reflect organization and requirements of the PRACE AISBL, which is the central entity for the research infrastructure and the first interface with users and scientific projects.

Also internal services needed for a effective collaboration among all members of the PRACE project, have been defined and implemented (i.e. the PRACE WIKI).

DEISA experience has also provided valuable input for driving the evolution of the PRACE infrastructure. Previous outcomes and reports have been fundamental for laying the ground of the current services layer and the fulfilment of emerged user's requirements.

1.2 Objectives

Main objective of this first year is the integration of Tier-0 centres into a single Pan-European infrastructure providing high-end computing services.

In order to achieve this goal, a list of secondary objectives has been defined and archived.

a. Organisation of the activities following a service-centric perspective and definition of role and responsibilities for each service category which has been identified (network, data, computing, AAA, users, monitoring and internal services).

b. Selection of a software catalogue to be deployed on Tier-0 systems.

c. Definition and use of a collaborative environment for supporting Tier-0 centres on installation and configuration of common services.

d. Setup of monitoring mechanisms for tracing and verifying the deployment activities on Tier-0 systems.

e. Compatibility check of the selected services for future integration with Tier-1 systems.

f. Standardize, where possible, the way to provide services to users and the internal maintenance procedures.

g. Collect and analyse requirements from users. Much attention has been devoted in getting users collaborate within internal technical activities to be sure to match their requirements.

h. Consolidate existing technologies to continue along the path set out during the preparatory phase to ensure the same user interfaces to access the PRACE systems.

i. Define technical requirements in order to permit PRACE-RI evolve in a consistent manner as new Petascale systems will be integrated into the PRACE-RI.

j. Set up a formalized process to guide the evaluation of new technologies and the design of new services.

k. Evaluate new technologies on the base of identified requirements.

l. Improve the quality level of available services.

1.3 Operational procedures and methodologies

Operational procedures and methodologies that have been adopted both for service deployment and for service assessment are out of scope for this first annual report and are included in the deliverable D6.1.

The PRACE Security Forum and the PRACE Operational and Coordination Team represent the two operational units that are responsible for taking decisions on security issues and for coordinating the software management on Tier-0 sites. Both are involved in Task 6.2 and Task 6.3.

The PRACE Service Catalogue and the Change Management and Incident procedure are also described in D6.1; they are an important reference for improving service deployment and assessment activities. In particular the PRACE Service Catalogue, referenced in this report, is directly linked with the Software Catalogue. The first mentioned defines services provided by PRACE-RI, the second mentioned defines how these services are implemented on Tier-0 systems.

The Change Management procedure defines the steps to be followed for a change on a deployed service or the integration of a new service that has been assessed and tested by Task 6.3.

All the work has been driven by a common methodology that is represented by a service-centric approach.

As described at the beginning of this introduction, all services are classified in six categories. For each category a responsible person is identified with the role to coordinate sub-activities related to deployment and evaluation. This person is also the first point of contact for Tier-0 sites for any kind of problem related to a specific service.

1.4 Structure of the document

The remainder of this document is divided into 5 chapters.

- Chapter 2 provides a technical overview of the Tier-0 systems that are currently in production.
- Chapter 3 summaries all activities carried out within Task 6.2 during the first year. All information is organized into sub-sections, each one reflecting the work done within each service category.
- Chapter 4 is dedicated to Task 6.3 and reports on assessment of new technologies following the same structure as Chapter 3.
- Chapter 5 is dedicated to internal services, which are used for both coordination and operational activities by members of the project.
- A final chapter “Conclusions and future work” links the present activity with the planned developments of the PRACE-1IP, which will be finalized in the second annual report.

2 Status and planning of Tier-0 services

This chapter provides a technical overview of PRACE Tier-0 systems, currently in production and available for the users:

- JUGENE, installed at GCS@FZJ
- CURIE, installed at GENCI@CEA

The next Tier-0 system (HERMIT, GCS@HLRS) will become available for the third call for access, currently open until June 22nd 2011. A tentative timeline to put in production other Tier-0 systems will also be presented at the end of this chapter.

2.1 Technical overview of the current Tier-0 production systems

The scope of this section is to provide a set of technical information about Tier-0 systems currently in production. Information about the application software environment is out of scope of this report and managed by PRACE-1IP-WP7. Detailed and dynamic data about deployment activities is part of the next chapter.

2.1.1 *JUGENE – GCS@FZJ*

The first German Tier-0 system JUGENE is managed by Jülich Supercomputing Centre. The water cooled, 72 rack, IBM BlueGene/P system has been installed in 2009 and was running in full production, when it was first offered to the PRACE community in the summer of 2010.

Figure 1: JUGENE at GCS@FZJ

JUGENE compute nodes are managed by a compute node kernel (CNK) and an I/O node kernel (Linux) runs on the 600 I/O nodes. A 3-dimensional torus network interconnects all compute nodes along with a collective and a global barrier network. The external connection is realized via 10Gb Functional Ethernet. Jobs are launched from two Front-End-Systems (Linux) to the IBM Tivoli Workload Scheduler LoadLeveler.

IBM's General Parallel Filesystem (GPFS) provides transparent access to scratch and home file systems managed by the central fileserver JUST (Jülich Storage Server), which also provides long-term storage and archiving to tapes via IBM Tivoli Storage Manager (TSM/HSM) (***Figure 2***).

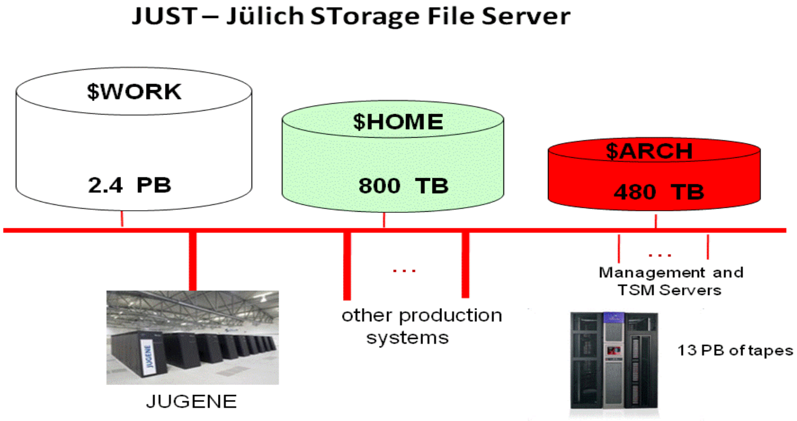


Figure 2: JUST storage architecture

Basic information

MACHINE NAME	JUGENE
PRACE partner	FZJ (GCS)
Country	Germany
Organisation	FZJ
Location	FZJ, Jülich, Germany
Nature (dedicated system, access to system, hybrid)	Access to production system
Vendor/integrator	IBM
Architecture	Blue Gene/P
CPU (vendor/type/clock speed)	IBM / PPC450d / 850 MHz
CPU cache sizes (L1, L2, L3)	32 KB per core / 8 MB shared
Number of nodes	73728
Number of cores	294912
Number of cores per node	4
Memory size per node	2 GB
Interconnect type / topology	Proprietary / 3D- torus + tree
Peak performance	1008 TFlop/s
Linpack performance (measured or expected)	825,5 TFlop/s measured
I/O sub system (type and size)	10 GbE connected GPFS server Approx. 5,6 PB
File systems (name, type)	/homeX, /work, /archX (all GPFS)
Date available for PRACE production runs	July 2010
Link to the site's system documentation	http://www.fz-juelich.de/jsc/jugene

Table 1: Basic information of JUGENE at GCS@FZJ

2.1.2 *CURIE – GENCI@CEA*

CURIE is the French PRACE Tier-0 system owned by GENCI and managed by CEA.

CURIE is provided by Bull and based on their bullx architecture. The supercomputer provides 2 types of computing resources each of them matching a delivery phase.

The first type of computing resources is based on the Bull Mesca fat node hardware, which has been installed at the end of 2010 and which is available to PRACE users since February 2011.

The second type of computing resources of CURIE will be based on the Bull INCA system based on thin node hardware, which will be deployed at the end of 2011 to be available to PRACE users in February 2012.

Figure 3: CURIE at GENCI@CEA

CURIE has been installed in the new computing centre of CEA, called TGCC, an infrastructure of 6500m² designed to accommodate future high end computing systems.

CURIE runs the bullx Linux 6 operating system Advance Edition, which delivers the Lustre file-system technology, the SLURM batch scheduler, the OFED InfiniBand software stack and the bullxmpi MPI stack.

TGCC center has been designed according to a data-centric architecture and provides a hierarchical data storage system that will manage 10 PB of data.

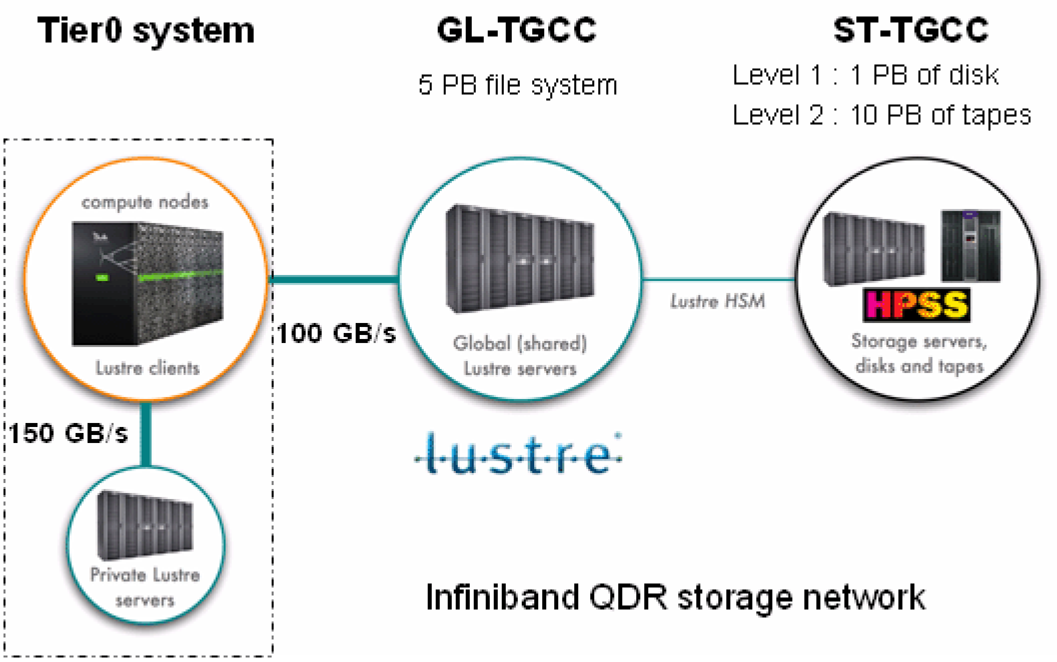


Figure 4: TGCC storage architecture

A global storage subsystem, called GL-TGCC, is used for data storage, handling and post-processing. GL-TGCC provides a global Lustre file system which is transparently connected with Lustre-HSM feature to a long-term storage and archiving subsystem. This subsystem is called ST-TGCC and uses IBM HPSS Hierarchical Storage Manager (***Figure 4***).

CURIE computing nodes are connected to a private Lustre storage system for very fast I/O. CURIE also uses Lustre routers to access the global Lustre file system through the backbone InfiniBand network.

Basic information

MACHINE NAME	**CURIE**
PRACE partner	GENCI
Country	France
Owner organisation	GENCI
Managing organisation	CEA
Location	CEA/DAM, Bruyères-le-Châtel
Vendor/integrator	Bull
Architecture	bullx
Phase 1	
CPU (vendor/type/clock speed)	Intel Xeon Nehalem EX 2.26GHz
Number of nodes	360
Number of cores	11520
Number of cores per node	32
Memory size per node	128 GB
Interconnect type / topology	InfiniBand QDR / FatTree
Peak performance	105 TFlops

Linpack performance (measured or expected)	*Not measured*
I/O sub system (type and size)	- Private storage over IB QDR, Lustre, 30 GB/s, 800TB - Global storage over routed IB QDR, Lustre, 10 GB/s, 400TB
File systems (name, type)	- /ccc/scratch, private Lustre storage - /ccc/work + /ccc/store, global Lustre storage
Date available for PRACE production runs	February 2011
Phase 2	
CPU (vendor/type/clock speed)	Intel Xeon SandyBridge
Number of nodes	5040
Number of cores	80640
Number of cores per node	8
Memory size per node	32 GB
Interconnect type / topology	InfiniBand QDR / Full FatTree
Peak performance	1.6 PFlops
Expected Linpack performance	1.2 PFlops
I/O sub system (type and size)	- Private storage over IB QDR, Lustre, 150 GB/s, 3.5PB - Global storage over routed-IB QDR, Lustre, 100 GB/s, 5PB
File systems (name, type)	- /ccc/scratch, private Lustre storage - /ccc/work + /ccc/store, global Lustre storage
Date available for PRACE production runs	February 2012
Link to the site's system documentation	http://www-hpc.cea.fr/en/complexe/tgcc-curie.htm

Table 2: Basic information of CURIE bullx supercomputer at GENCI@CEA

2.2 Planning and integration of new Tier-0 production systems

At the time of writing, technical information about future Tier-0 systems is not complete. The following table only summaries estimated timeframes for the service deployment process on future Tier-0 systems.

Machine Name	Organisation	Country	Ready for service deployment	Ready for production runs
HERMIT	GCS@HLRS	Germany	Q3/2011	Nov/2011
SuperMUC	GCS@LRZ	Germany	Q2/2012	Jul/2012
N/A	CINECA	Italy	Q4/2011	Q1/2012
N/A	BSC	Spain	N/A	2013

Table 3: Estimated timeframes for PRACE service deployment on future Tier-0 systems

3 Selection and deployment of common services

The process of selection and deployment of a common set of services aims at presenting all Tier-0 centres as a single distributed infrastructure, instead of a set of individual systems/computing facilities. Coordinating this process is the main objective of Task 6.2.

In general, a software deployment process includes different activities such as software installation, configuration and maintenance. Regardless of the actual workflow of activities, all of them are logged and documented in order to assure a high quality in service provisioning and to enhance a knowledge base among all partners.

Task 6.3, which is responsible for technology watch and assessment, has the role of being the main input of this process by triggering the introduction of new services/software to be brought into production on Tier-0 systems.

At the beginning of the project, results from the PRACE Preparatory Phase and DEISA have constituted a valuable starting point for releasing a first software catalogue.

From an operational point of view, the deployment process is almost fully defined. An additional procedure for Change Management will complete this process by providing a list of actions to be taken when a new service has to be deployed and/or an existing service has to be updated.

Each service category has a responsible person who is in charge of managing all the information and decisions related to a specific service area as well as supporting Tier-0 sites on operation activities.

PRACE WIKI is the main collaborative tool used to coordinate all activities undertaken by the Task 6.2.

The integration between JUGENE and CURIE, the two currently available Tier-0 production systems, has been the objective of this year. Other activities focused on the definition of common deployment procedures for future Tier-0 systems. For the next year, the focus will move to the integration between Tier-0 and Tier-1 computing infrastructures. In order to be well prepared, all decisions on software selection, which have been taken this first year, have already taken into account the future integration with the Tier-1 computing layer.

3.1 Overview of common services deployment

The first action has been the definition of a common software catalogue, a single point of reference for all Tier-0 centres and the result of the service selection process.

The software stack proposed in the preparatory phase has been reviewed considering the difference between the computing facilities that were considered, from system prototypes (even if representatives of Tier-0 systems) to high-end computing platforms.

For network services, Iperf [31] has been selected as the common tool for monitoring the status of the PRACE network, which relies on the GÉANT academic network provider [27]. Special attention is dedicated for providing different views of the complete dataset, which can be used both for service tuning and to improve network access. A dedicated web page is maintained to provide this information.

Data movement between PRACE Tier-0 systems can take advantage of the dedicated network links. GridFTP [43] is the selected software to allow users to move large data sets to, from, and within the PRACE infrastructure in a reliable way.

UNICORE [20] is the selected software for computing services provisioning. A configuration has been defined to offer users with basic job management mechanisms through three different user interfaces. UNICORE also provides a storage management services for user data staging on remote resources by supporting protocols such as HTTPS, FTP and GridFTP.

Accounting services run through a mutual and coordinated interaction between the central LDAP-based user management system of the PRACE-RI, where user profiles are stored, and a set of local databases, managed by Tier-0 centres, where usage records are collected. The DART [13] client can be used to retrieve accounting information and provide different levels of details in the basis of defined authorization policies.

Modules framework [37] is used to set up and provide users with a common production environment, in terms of default choice of compilers, tools, and applications, on each Tier-0 system.

Service monitoring is implemented in PRACE by using the Inca framework [14]. System administrators and users can check if all services provided by Tier-0 systems are correctly implemented and available.

In order to properly authenticate users and services on Tier-0 systems, an X.509 PKI [6] is required. To establish trust relations, the entire X.509 infrastructure relies on Certification Authorities (i.e., those Policy Management Authorities [5] federated with the IGTF [4]) that any involved party can trust to provide identity assertions. For proper and correct operation of the authentication infrastructure, all Tier-0 systems accept the same subset of Certification Authorities.

Following table lists all software selected during this first year with the deployment status on Tier-0 systems.

Service Category	Software	JUGENE	CURIE
Network	Iperf	Installed	Installed
Compute	UNICORE	Installed (v6.3.0)	Installed (v6.3.0)
Data	GridFTP	Installed (v3.23)	Installed (v3.28)
AAA	Central LDAP	Synchronised	Synchronised
AAA	PRACE Accounting facility	Installed	Available by Nov/2011
AAA	GSI-SSH (X.509-based authentication)	Installed	Installed
Monitoring	Inca	Installed (v2.5)	Available by Nov/2011
User	PCPE	Installed	Installed

Table 4: Software Catalogue and deployment overview

3.2 Network services

The PRACE network services are based on the developments done in the DEISA, eDEISA and DEISA2 projects. The Tier-0 system JUGENE at FZJ has been connected to the DEISA backbone by using the already existing connection of the Tier-1 Jülich JUMP systems, now sharing the 10 Gb/s link to Frankfurt with the current FZJ Tier-1 system JUROPA. The connectivity to JUGENE is up and running. Monitoring of the FZJ connection to the DEISA/PRACE backbone is done via the established production network monitoring.

At the end of last year the connection to the Tier-0 system CURIE at CEA has also been set up. CEA is sharing the 10 Gb/s link to the DEISA/PRACE backbone with CEA R&D computing centre and IDRIS.

In the future, an additional network setup will be done for the upcoming Tier-0 systems at HLRS Stuttgart, sharing the link with the Tier-1 system, and LRZ Garching near Munich, sharing the link with Tier-1 systems at LRZ and RZG.

Further systems will be added when available.

3.3 Data services

GridFTP is the de-facto standard for bulk data transfers. It provides high performance data transfers and is able to exploit the high-speed interconnect between the current Tier-0 machines CURIE (GENCI@CEA) and JUGENE (GCS@FZJ).

In 2009 PRACE-PP also suggested the use of RFT (Reliable File Transfer) as data service built on top of GridFTP, but the Globus Alliance dropped support for RFT since Globus Toolkit (GT) 5. Consequently, data services in PRACE-1IP are currently based solely on the Globus GridFTP software. This is not a step backward, as RFT only offered a few additional features. Additionally the migration to GT 5 is also mandatory because of security reasons, as older GT versions are no longer supported by the Globus Alliance.

The DEISA project has been using GridFTP for a long time and has built up a lot of expertise with this service. It was only natural to benefit from this expertise. Therefore DEISA resources about GridFTP were collected and reviewed for useful information for PRACE-1IP. This included not only documentation but also specific software developed in DEISA that could also be useful for PRACE-1IP. For example in DEISA an installation and setup 'how-to' document for the GridFTP software was created. This was adapted for PRACE-1IP and was recently moved to the PRACE WIKI facility. Other examples are the init scripts for easy integration of the GridFTP service in System V (SysV) based init systems or specific tools that can ease the set-up of data transfers for the user.

During the first year of PRACE-1IP the following additional items were created:

- An init script for Scientific Linux 6 (a variant of Red Hat Enterprise Linux 6 - RHEL6) was developed which offers the same functionality as other existing scripts. PRACE-1IP can now support SUSE Linux Enterprise Server 10 (SLES10) (support for SLES11 will be evaluated in the future), Scientific Linux 6 (SL6) (which should also give support for RHEL6) and Ubuntu 10.04 LTS. To support all common operating systems a version for AIX needs to be created in the future. All software will be made available on the PRACE SVN service.
- A proposed setup document for the GridFTP service was developed. It gives specific suggestions for the deployment taking into account both security and performance. It also gives hints for interfacing to the backend storage and makes it possible to serve all supercomputer systems at a site. References to the installation and setup documentation and the SysV init scripts complete the document. With this information a hosting site is able to implement a GridFTP service with minimal efforts. The documentation is available in the PRACE WIKI facility.
- A proposal for GridFTP monitoring for both performance and availability was created and is now entering the evaluation phase.

Combining all these parts will result in a sustainable data service for the users.

At the time of writing the current GridFTP infrastructure in PRACE-1IP includes two Tier-0 systems. Detailed information about the service implementation will be collected in the GridFTP inventory in the PRACE WIKI facility.:

a) CURIE (GENCI-CEA): The GridFTP software is already deployed at CEA and the service is currently tested internally. After internal tests are successful, the service will be made available to users.

   The planned configuration will include six hosts responsible for data transfers and one virtualized host responsible for control communication. GridFTP v3.28 (GT 5.0.3) is used for the service.

b) *JUGENE* (GCS@FZJ): FZJ currently provides GridFTP access to JUGENE through a GridFTP service on their JuRoPa machine as both machines share file systems. This is only an intermediate solution, because there will also be a dedicated GridFTP service made available on JUGENE in the future. This service is already using the SysV init script for SLES10. After final testing, this service will be made available to users.

   The current configuration includes one host both for data transfers and control communication. GridFTP v3.23 (GT 5.0.2) is used for the service.

By using the high speed interconnect between the current hosting sites, the Tier-0 machines are also connected to the sites (and Tier-1 machines) that participated in DEISA, which enables data transfers from external sites or institutions.

The next Tier-0 system will be *Hermit* (GCS@HLRS), which is expected to be available in November 2011. HLRS plans to use at least one machine dedicated to GridFTP and to run the current stable GridFTP version. Deployment will be made in accordance with the proposed setup.

The following is a list of planned activities for the next year:

- *Deployment of advanced user tools for data transfer.* This will include a comparison of two advanced tools developed in DEISA and the decision for one tool to deploy and to promote.
- *Creation of an inventory of deployed GridFTP services.* This inventory will include detailed information about each service and will help in diagnosing potential problems.
- *Monitoring data services.*
- *Risk review of GridFTP service.* It is part of the ordinary software maintenance activity and it aims to foster security of the GridFTP service in PRACE-1IP. A risk review is planned to identify possible (security) risks and countermeasures.
- *Service certification of GridFTP service.* Please see paragraph 4.4 for details.

3.4 Compute services

Primary goal of this activity is to deploy and maintain solutions enabling compute services.

In this first year, the main activity is focused on setting up a common deployment procedure for UNICORE v.6 on Tier-0 systems. In particular:

1. The definition of a common configuration and a common set of components provided by UNICORE;
2. The creation of a document repository for supporting Tier-0 system administrators;
3. The selection and implementation of the central services provided by UNICORE;

4. The collection of post-installation feedback from Tier-0 centres.

UNICORE services can be divided in central services (Workflow Orchestrators, Registry, CIS), and site-specific services (GATEWAY, UNICORE/X, XUUDB, TSI).

For site-specific services a comprehensive documentation has been provided by FZJ; it is available for download from the PRACE collaborative workspace (PRACE WIKI). The documentation is available for version 6.3, which is the official production release used by PRACE.

As central services, only the “Registry” component has been considered as mandatory and to be deployed on Tier-0 systems. The adoption of other central services, like the workflow engine, is currently under assessment. The UNICORE Registry is necessary to build up and operate a distributed infrastructure. It is installed at FZJ, while a backup server is available at CINECA to prevent a single point of failure in the infrastructure.

All technical aspects about installation and configuration are available at PRACE WIKI.

A test suite for UNICORE is currently under definition. This activity should encourage system administrators at each site to produce feedback about installation and operation of UNICORE.

While UNICORE represents a solution for providing a common interaction layer to different computing facilities, local batch scheduler systems are always the way of a direct interaction. This is especially true within a HPC ecosystem, like the PRACE Tier-0 infrastructure, where users are generally interested to use the computational power of a single high-end system.

A common repository has been created for collecting all different features provided by local batch systems running at Tier-0 level. This set of information can be used by WP7, which is working on user level, but also by other internal activities like those carried on by the PRACE security forum and/or software customizations (i.e., customizations for UNICORE TSI).

3.4.1 *Deployment of compute services on Tier-0 Systems*

UNICORE basic components are currently installed and available on both Tier-0 production systems on the public Internet. The supported version is UNCORE 6.3.2.

JUGENE system also offers support for managing workflow of jobs, which is an optional feature so far.

The following table summarises the deployment status of UNICORE on JUGENE and CURIE.

Site	System	Network	Version	U6 GATEWAY	U6 UNICORE/X	U6 XUUDB	U6 TSI	U6 WF
FZJ	JUGENE	Public	6.3.2	Yes	Yes	Yes	Yes	Yes
CEA	CURIE	Public	6.3.2	Yes	Yes	yes	Yes	No

Table 5: Status of UNICORE on Tier-0 systems

Even if the deployment of Batch Scheduler Systems is a local operational activity at each site, a configuration inventory is kept and maintained by this task.

Following table summaries the BSS currently deployed. List of keywords and features used for each BSS are collected in the PRACE WIKI.

Site	System	BSS	Arch	OS
FZJ	JUGENE	LoadLeveler	BlueGene	Linux
CEA	CURIE	SLURM	x86_64	Bullx Linux 6

Table 6: Batch Scheduler Systems on Tier-0 systems

3.5 AAA services

The AAA activity is responsible for services which provide Authentication, Authorization and Accounting facilities on the infrastructure. This includes the provision of interactive access, the authorization for services and the provision of information on the usage of the resources.

The implemented facilities are based on the evaluation results of the PRACE-PP project and most are based on the solutions developed for the DEISA infrastructure. The use of DEISA developed facilities will enable a smooth integration of the Tier-1 infrastructure, the continuation of the DEISA infrastructure.

3.5.1 *Public Key Infrastructure - PKI*

Several PRACE services rely on X.509 certificates [6] for the authentication and authorization. These certificates must be issued by entities which are trusted by the service providers, meaning among others that the attributes tied to a public key are properly validated and that the owner of the attributes (a person) or the responsible entity for the attributes (of a server) can be traced.

PRACE relies on the Certificate Authorities (CA) accredited as a member by the EUGridPMA, the European Policy Management Authority [5], or by one of the two sister organizations TAGPMA and APGridPMA, all three federated in the IGTF [4]. These PMAs all require a minimum set of requirements for the CP/CPS of the member CAs, published in a profile document.

Different operational models for CAs exist: Classic X.509 CAs, Short-Lived Credential Services (SLCS) and Member Integrated Credential Services (MICS) and all three have their own profile, which is published on the IGTF website under the header Authentication Profiles [4]. PRACE is a member of EUGridPMA as Relying Party (RP), which gives the opportunity to provide feedback on internal needs and also to monitor the accreditation of new members and the audits of existing members. The 22nd EUGridPMA meeting from 11-13 May 2011 in Prague was attended by a PRACE representative and a presentation was given with an introduction of the PRACE project and a discussion of some of internal issues in the area of authentication and authorization.

Relying on these CAs is also important for the collaboration between infrastructures, because other infrastructures like EGI, TeraGrid also trust these CAs. A user can use the same certificate for authentication and authorization purposes on different infrastructures. Of course the authorization decisions remain the responsibility of the infrastructure to which the user authenticates.

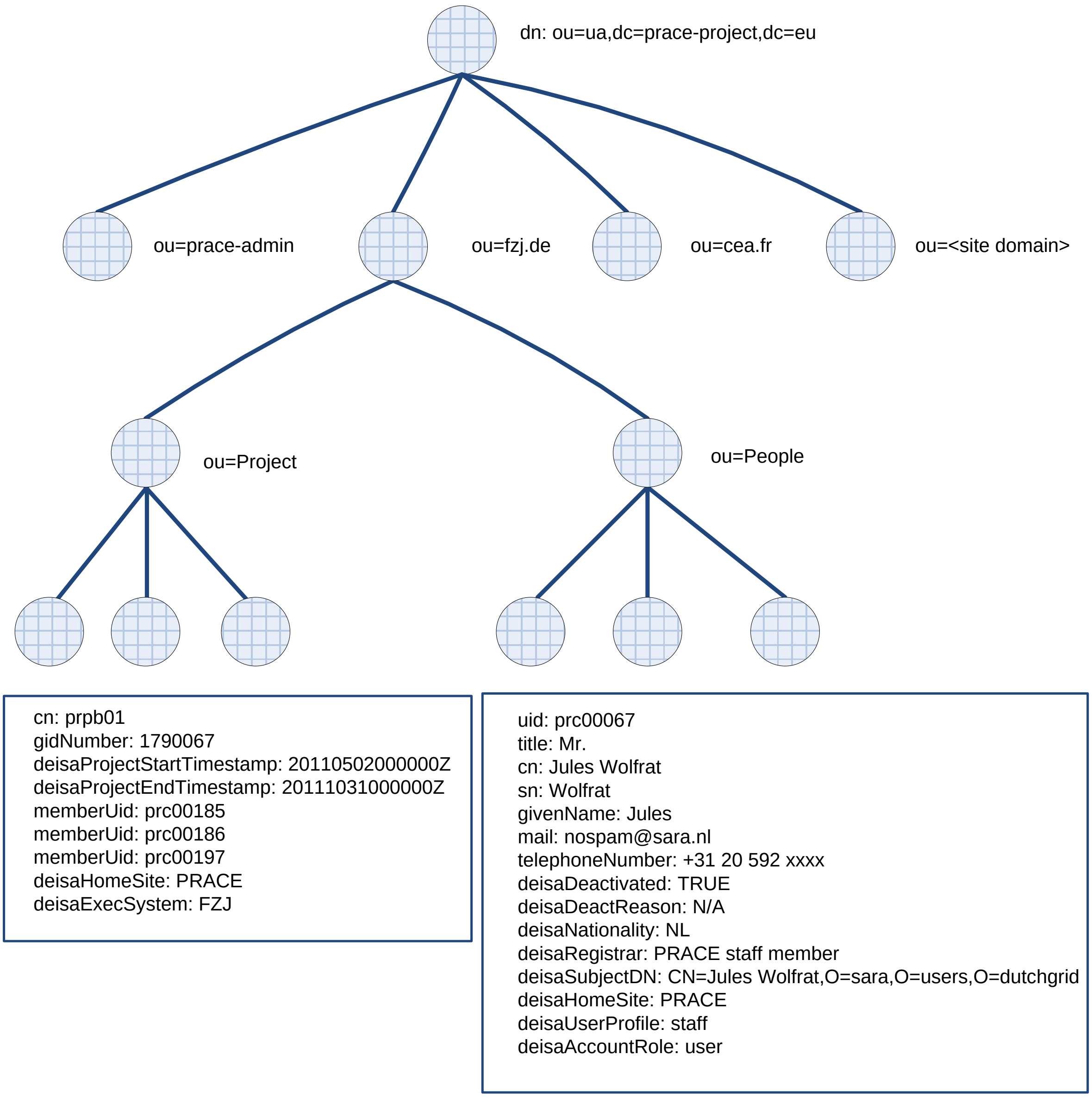
For PRACE a distribution of CA information is maintained at a central repository [7]. This distribution is shared with DEISA because the same requirements exist for both infrastructures. The distribution is provided in several formats because services have different requirements for the presentation of the information. When the IGTF distribution is updated (several times a year), the PRACE/DEISA distribution is also updated and all partners are required to update the local repositories used by services to validate the authentication and authorization of users and other services.

The possibility to revoke certificates is an important property of a PKI. CAs must publish a list of X.509 certificates, CRLs (Certificate Revocation Lists), which should not be trusted anymore. Services must use these CRLs as part of the authorization decision. Also certificates have a specific validity period, but this is published in the certificate itself. The primary use of the CRL is for certificates that must not be trusted anymore before the validity has expired. For instance, if it is known that the private key of a certificate is stolen the certificate must be revoked. Partners can use a script, fetch-crl, for the periodic retrieval of CRL information [8].

3.5.2 *User Administration*

Information about users and their accounts is maintained in an LDAP-based repository. This facility is used to update the authorization information needed by services and in general can be used to retrieve information about users and the projects that they are affiliated to. Authorization information is provided among others for interactive access through GSI-SSH, job submission with UNICORE, accounting services and access to the helpdesk facilities.

One LDAP server is used for PRACE, operational since July 2010 at SARA. The schema used to describe the information is the same as used by DEISA. This enables a smooth integration of the Tier-0 infrastructure with the Tier-1 infrastructure. For instance the same tools can be used to update and retrieve information from the different repositories. However, the namespaces differ so that it is clear to which domain the information belongs. ***Figure 5*** shows the LDAP schema for PRACE together with examples of a project entry for prpb01 and an account with uid prc00067. The difference with DEISA repositories is in the top level domain, where *dc=prace-project* is replaced by *dc=deisa* for the DEISA namespace.



```
cn: prpb01
gidNumber: 1790067
deisaProjectStartTimestamp: 20110502000000Z
deisaProjectEndTimestamp: 20111031000000Z
memberUid: prc00185
memberUid: prc00186
memberUid: prc00197
deisaHomeSite: PRACE
deisaExecSystem: FZJ
```

```
uid: prc00067
title: Mr.
cn: Jules Wolfrat
sn: Wolfrat
givenName: Jules
mail: nospam@sara.nl
telephoneNumber: +31 20 592 xxxx
deisaDeactivated: TRUE
deisaDeactReason: N/A
deisaNationality: NL
deisaRegistrar: PRACE staff member
deisaSubjectDN: CN=Jules Wolfrat,O=sara,O=users,O=dutchgrid
deisaHomeSite: PRACE
deisaUserProfile: staff
deisaAccountRole: user
```

Figure 5: PRACE LDAP namespace

As can be seen from the examples several attributes have DEISA in their name, which only means that the DEISA defined schema is used. These attributes were defined to provide information which could not be provided by standard schema definitions.

Because the repository contains private and sensitive data the security of the facility is important. To insure confidentiality, privacy, message (data) integrity and non-repudiation TLS is used for all communication. Only staff members of PRACE partners and selected servers at partner sites have access to the service. The authentication and authorization is based on X.509 certificates. Write access to the repository is further limited, persons are only granted write access for the domain(s) for which they are responsible. The division in different site domains enables a fine grained access control, e.g. only staff of FZJ only can update entries under the *ou=fzj.de* domain (see ***Figure 5***).

Information about projects is maintained in the database used for the peer review of project applications (PRACE Peer-Review Tool). This information is copied into the LDAP repository to enable easy access to the data. ***Figure 6*** gives an overview of the data flow to and from LDAP. After a project proposal is accepted, CINES, the partner which manages the peer review database, registers the projects in LDAP and informs the Tier-0 sites about the users that should be given access. The Tier-0 sites start the administrative tasks required to complete the registration of the users in LDAP and on their systems. Once all information is in LDAP all services can use the information to update the authorizations, e.g. for the trouble ticket system, accounting services etc.

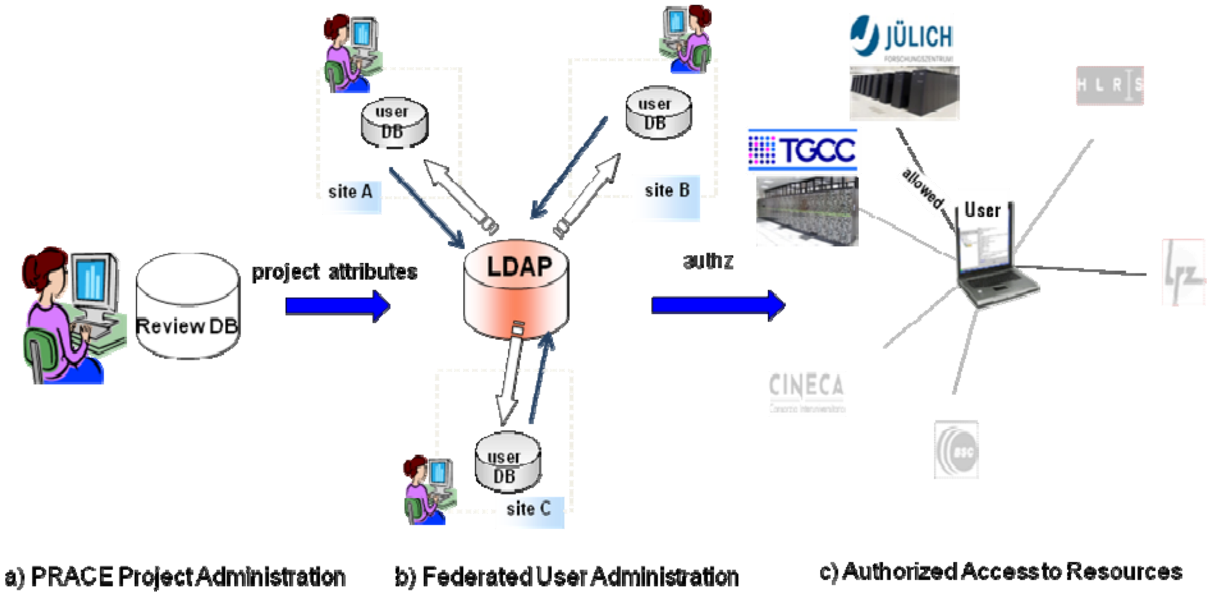


Figure 6: Overview of project and user administration with LDAP

Also staff members can get access to the Tier-0 services. WP7 members can get accounts for application enabling or benchmark tests and WP6 members can get access for the testing of services. For both current Tier-0 systems, JUGENE and CURIE, resources are allocated for the support functions provided by WP7 and WP6. The WP7 allocations are divided again between the different tasks of WP7. Accounts are created by the Tier-0 sites after approval by the WP7 task leaders or the WP6 leader. These accounts are also registered in the LDAP and procedures for the registration are defined. A different administrative domain, prace-admin, is created, so that the responsibility for this domain can be separated from the responsibilities for standard user accounts.

The status at the end of the first project year is that users for both Tier-0 production systems and other PRACE services are registered in the LDAP, and that authorizations for the services are provided by the LDAP. Several video conferences were organized to discuss the details of this service and the procedures and policies agreed for the management of the user administration will be documented in what will become the “PRACE AAA - Administration Guide”.

3.5.3 *Interactive access*

Interactive access to the Tier-0 systems is a basic requirement. This is provided using the SSH (Secure Shell) facilities provided by most distributions of operating systems. However, the standard distribution of the popular OpenSSH [9] implementation does not support X.509 certificates for authentication and encryption. The Globus community distributes a X.509 based OpenSSH version, GSI-OpenSSH [10] or GSI-SSH for short. This tool is accepted by PRACE to provide interactive access to systems. On JUGENE and CURIE, GSI-SSH based access is enabled. GSI-SSH clients are available too, one of these, GSI-SSH_Term, is supported by PRACE partner LRZ.

3.5.4 *Accounting services*

Information about the usage of resources is important for users, Principal Investigators (PIs), partner staff and the management of the resources. The facilities developed within DEISA are accepted by PRACE. Important characteristics of the facilities are: 1) the usage of resources is published in a common format, following the recommendations of OGF's UR-WG (Usage Record Working Group) [11]; 2) authorized access based on the rights of the requestor, e.g. a normal user can only see his/her own personal usage while the principal investigator of a project can have a more detailed information set. Detailed information about the design considerations can be found in [12].

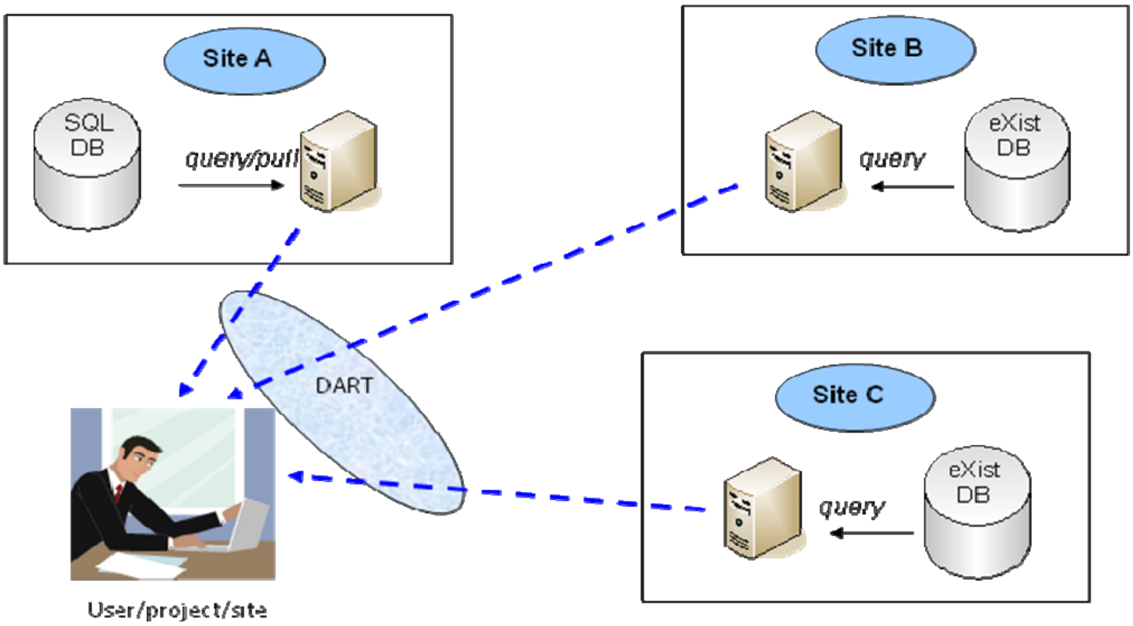


Figure 7: Accounting architecture

Figure 7 shows the basic set-up of the facilities. Each site stores usage records for PRACE users in a local database, this can be a specific eXist database (sites B and C in the figure) or an SQL based database (site A). For each job the following information is stored: the system the job has run on, job identification, system wall clock time, CPU time used, number of cores used, project name, user identity (uid and X.509 subject DN) and the time that the job was submitted, started and finished. An Apache/CGI web interface is available which will provide data to a client if authorized. The authorization is based on X.509 certificates and the access rights are given by the attribute deisaAccountRole of the user administration service. DART [13] is a Java Webstart tool, which can be used by a client to retrieve and to display the information from different sites.

For JUGENE, the Tier-0 system at FZJ, the facilities are implemented, while for CURIE, the Tier-0 system at CEA, the facilities are planned to be available at the start of the third regular call, November 2011. On the CURIE system usage information is already provided by local facilities. In the current draft of the contributor's agreement between the PRACE-RI and the Tier-0 site is a list of reports that the site must deliver to the PRACE-RI:

1. Total usage relative to total available capacity (CPU hours);
2. Total usage relative to allocation per project;
3. Total usage per discipline;
4. Total usage per country of the PI's institution;

5. Distribution per job size and job duration.

Not all of this data is available through the PRACE accounting facilities yet and Tier-0 sites have to provide this data using local information. Using additional information from the user administration service it should be possible to provide most of these reports. Further development of reporting tools is required to include this information.

3.6 User services

3.6.1 *PRACE Common Production Environment*

The concept of a *Common Production Environment* is both straightforward to understand and commonly used today on distributed infrastructures. From the user's point of view, there is a clear necessity to be able to access a coherent set of software on the different sites of a distributed infrastructure, and through a common interface.

The persons in charge of such infrastructure have to define this set of software and to provide the common user interface on their own platform. This concept was already defined and used at the early stage of the American TeraGrid project [21], nearly ten years ago, at this time implemented as part of the *Coordinated TeraGrid Software and Services* (CTSS) [34] bundle.

In the same direction, and for the same obvious purposes, the *Distributed European Infrastructure for Supercomputing Applications* (DEISA) project defined and implemented this kind of framework since its beginning in 2004. It was named the *DEISA Common Production Environment* (DCPE)[1].

Using the experience achieved in the DEISA (May 2004 to April 2008) and DEISA2 (May 2009 to April 2011) projects, the initial PRACE-PP project (January 2008 to June 2010), in its Work Package 4 (*Distributed System Management*) has strictly adopted the same concept and framework as in DEISA, naming it the *PRACE Common Production Environment* (PCPE), which was pursued in PRACE-1IP WP6 and integrated as part of the *User Services* to provide to the PRACE users. Up to now this has only applied to Tier-0 site users, but it is planned to be extended to Tier-1 site users in the upcoming PRACE-2IP project.

The PCPE offers a common interface to the users, independent of the target platform employed. The level of coherence is not the same everywhere in the infrastructure, ranging from very high inside each subgroup of homogeneous computers (when several similar supercomputers will be later integrated as Tier-1 machines) to a lower level for the other platforms, as this is obviously the case for the Tier-0 machines which will be, at least in the short term future, all different.

The two components of the PCPE are:

- a coherent set of software packages divided into five categories: shells, compilers, libraries, tools and applications (as explained in a previous section, not the same software are available on all platforms, according to their status of *optional* or *additional* software, outside the *core* ones required to be available on each platform);
- a uniform interface to access the software, provided by the *Modules* tool[2] [37] originally developed by SUN but available today as a public domain software (in

[1] Lately, TeraGrid has introduced a more specific concept than the CTSS one, focusing only, on the software available and on their user interface, named the TeraGrid *Common User Environment* (CUE) [36], very analogous to what the *DEISA Common Production Environment* was.

[2] Initially, TeraGrid used the *SoftEnv* tool [38], developed at the Argonne National Laboratory, but later it introduced also the availability of the *Modules* tool and interface, due its growing popularity and usage.

several different implementations). This defines the *PRACE Modules Environment* (PME).

In this framework, each component of the Software Stack available in the PCPE of one computer is accessible by using a dedicated interface based on Modules and called a *modulefile*. For each piece of software, a corresponding *modulefile* must be present, which will be internally different to hide the specific characteristics of the installation of this software on each platform (especially the directories in which its components are physically installed, which often vary from computer to computer).

During the development of these *modulefiles*, special care has been taken to make the *Modules* commands as analogous among heterogeneous computers.

The internal implementation of them on the various computers can be different but the user interface is the same which allows to keep a high coherency between the computers of the heterogeneous infrastructure, offering the users a unique common interface on all platforms and allowing to describe only one interface for all the platforms in the user documentation.

Additionally, as long as that the user interface is kept strictly equivalent, each partner is free in the way they implement it on their computers.

It is especially beneficial to allow the sites to implement it as a separate independent environment available only for PRACE users or included in the generic Modules environment provided to all users.

The usage of this framework has been defined to be as simple and straightforward as possible for the users. However, this requires some additional complexity and effort in its development.

In particular, the major well known weakness of the Modules tool is that it considers each piece of software as an independent component without management of the possible dependencies between them (like between a dedicated implementation of the MPI library and a parallel version of numerical library which rely on it).

The robustness of the system has also been emphasized, with special checks and coherency tests, in order to prevent users from defining incompatible choices that would later create unexpected problems that may be difficult to diagnose.

To help partners and to decrease the amount of work required for them to implement the PME, a set of templates (one developed for the IBM Power systems and another one for the IBM BlueGene systems) has been provided by IDRIS, allowing others to implement the PME on their own system in an easier way.

These templates were in fact developed for the DEISA Modules Environment (DME), but as discussed and agreed two years ago between the WP4 PRACE Preparatory Phase and the WP6 DEISA2 work packages, the DME has been made at this time fully compatible with what was expected to be deployed on the future PRACE systems.

3.6.2 *Status of PCPE on current Tier-0 Systems*

The PCPE is deployed on the two current Tier-0 systems CURIE at CEA and JUGENE at FZJ. It is loaded using:

```
module load prace
```

Loading the "*prace*" module does not erase any local Modules environment as "*prace*" is a module like the others. Additional site specific modules may be loaded at each site, but are lower in the module hierarchy than the "*prace*" modules.

Dedicated monitoring of the PCPE will be implemented, in the same way as in DEISA for the DCPE. This monitoring is based on the Inca [14] monitoring tool initially developed for the TeraGrid project. This tool will provide a global view of the status of the Software Stacks on each of the systems within the distributed infrastructure. Further details of the implementation of the PRACE monitoring services are included in paragraph 3.7.

3.6.3 *User Documentation*

The availability of user documentation is of course an absolute requirement. However, as during the first period of the project, until May, only one platform at FZJ was available to the users, who could easily rely on the local documentation available at this site. A second Tier-0 site CEA was operational starting from May, and the third Tier-0 site will be online at HLRS in autumn. PRACE-2IP will be starting shortly, and we will have Tier-1 sites later in the year.

Common documentation will be required to support this, and will be provided as part of the PRACE User Documentation. Its expected content will be both a generic PRACE Primer, describing, without any technical details, the different services available, and dedicated independent user manuals containing for each of these services all the technical needed by users.

A large part of the technical information is already written and internally available using updated information from the corresponding DEISA user manuals [39], in particular for interactive access to the resources, file transfers, usage of the PCPE (see the previous section), usage of UNICORE 6.

These documents are not yet publicly available, because some technical choices, mostly related to the necessity to have a smooth integration of the on-line versions of these documents inside the new PRACE-RI Web site, which uses a different technology than the previous PRACE Project Web site, have not yet been finalised.

3.6.4 *The PRACE Trouble Ticket System*

A centralised Helpdesk is vital if we are to present PRACE as a single distributed architecture to all users. Contributors are required to provide local support in line with the PRACE Contributors Agreement, however a central Trouble Ticket System is required so that a single common view of all user issues can be maintained.

In PRACE-1IP we have selected to follow the model which was proven in DEISA. The DEISA Trouble Ticket System (TTS) was based around Best Practical Request Tracker (RT) Software, Version 3.6. A number of customizations were applied to the RT configuration during the DEISA project which are relevant also for PRACE also. In WP6 we have chosen to capitalize on the work of DEISA, but to also move the TTS to a more stable environment. A new installation of the RT has now been installed at CINECA.

The PRACE TTS utilises Best Practical Request Tracker V3.8. There are two instances available – one for test and development and one production instance. These use a Debian 6 platform, and are deployed on a Virtual Server Infrastructure. The Debian packaging enables straightforward upgrades of the software as and when required.

First line of support for the TTS is provided by CINECA. Second line of support is also provided by CINECA in order to support any configuration changes, or development of additional features as deemed necessary at a later date. The status of the TTS is monitored and backups are available which can be restored in the event of any major failure. Work is currently underway within WP6 to investigate options for providing a failover instance of the

TTS at a secondary site. This would involve the setup and configuration of a secondary virtual machine, with a synchronisation process in order to keep the failover instance up to date.

Access to the PRACE TTS is controlled by X.509 trusted certificates.

Both end users and support staff have access to the TTS. Access is authenticated using the data in the PRACE LDAP. Configuration of the support queues within the TTS is currently in process.

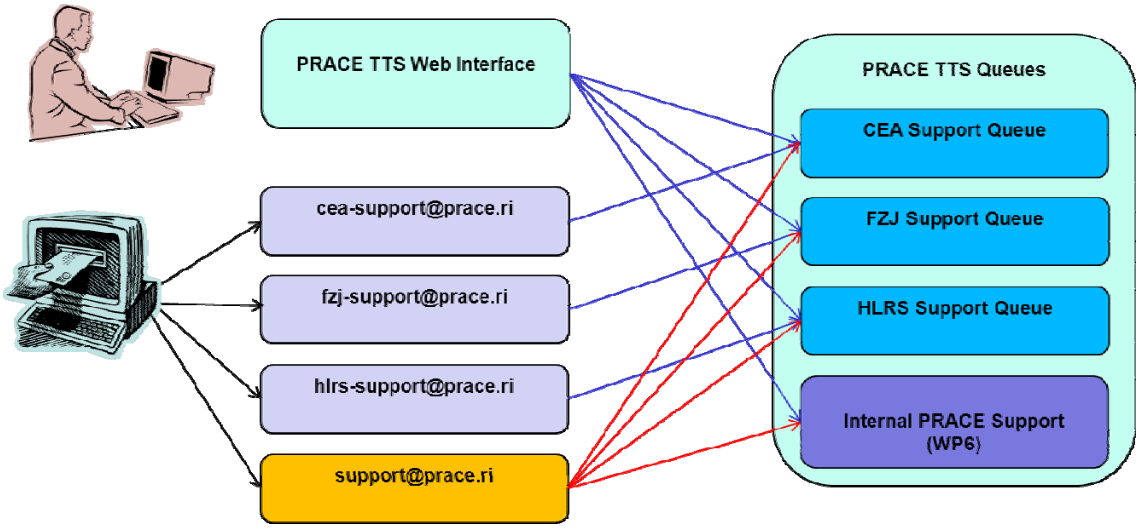


Figure 8: PRACE TTS Configuration

The primary interface for users is via the TTS web interface. This enables users to select criteria which routes their issue directly to the Contributor providing support for the system which they have been allocated resources on. Issues raised in this manner would be routed automatically with no manual intervention.

At the time of writing, a secondary email-based support has been defined with minor implementation aspects to be solved.

This option would be required if a user was unable to access the web interface for any reason (such as a problem with their X.509 certificate). A generic email address (support@prace-ri.eu) would be routed to the PRACE TTS. Issues raised in this manner would be monitored by a duty Helpdesk team, who would in turn route the issue to the appropriate Contributor.

In response to concerns raised, the ability to email site specific email addresses will also be configured, e.g. cea-support@prace-ri.eu. Issues raised via this email route would not require manual intervention by a duty Helpdesk team, but would also route automatically to the Contributor. A concern here is that this approach does not present PRACE as a single, distributed research infrastructure to users, but instead presents PRACE as a collection of independent sites. Users will be encouraged in all documentation to use the TTS web interface whenever possible.

Additional internal TTS queues are available for support staff only (i.e. those staff providing support at a Contributing site or WP6 internal staff), which enables tickets to be routed internally to those teams providing support for the distributed services on the PRACE-RI such as network, AAA, monitoring etc. Tickets raised by end users would not be routed automatically to these queues. The Contributing site would always provide the first line support, but would escalate internally to these queues if the issue was with a centrally installed component of the distributed infrastructure.

3.6.5 *Visualisation Services*

Visualisation services are very important today for the analysis of the results produced by simulations. The files generated as a result of the simulations run on today's supercomputers are often very large and are therefore difficult to transfer or post process at a user's local site. The requirement to be able to access Remote Visualisation Services (RVS) has arisen as a result.

During the DEISA2 project an RVS was deployed as part of Work Package 4 (led by CINECA). The DEISA Remote Visualisation Service was provided by LRZ. It is not expected to install any visualisation services at the PRACE Tier-0 sites but to utilise the existing LRZ facility.

Full details on the service provided are available in the LRZ Remote Visualisation Guide [40].

The complete list of applications is available at [41].

3.7 Monitoring services

The goal of the monitoring services sub-task is to deploy and operate solutions and tools for monitoring of the availability and functionality of the PRACE e-Infrastructure components and services. A variety of applications geared towards particular monitoring scenarios is currently available on the market. For example, applications designed for monitoring of hardware status, service functionality or computing environment availability through different patterns and offering different feature set.

Taking into account that HPC systems and the underlying infrastructure is monitored internally by the respective PRACE partners in this sub-task we primarily focus on monitoring of availability and functionality of the PRACE services from the user perspective. For this a user-level monitoring Inca [14] is used.

Inca is an application for monitoring of a computing infrastructure from a user point of view. It is developed by the San Diego Supercomputing Center and is used worldwide by many distributed computing infrastructure projects. In 2005 Inca was chosen by the DEISA project for monitoring the DEISA Common Production Environment (DCPE). Later its usage was extended to DEISA Middleware services, the DEISA GPFS and the DEISA LDAP user administration system. Based on the DEISA recommendations this application is deployed in PRACE for monitoring of the e-Infrastructure components and services offered to the users.

Inca implements the client-server model where Inca clients called reporter managers are testing components of the PRACE e-Infrastructure and sending the collected monitoring data to the central Inca server for processing, archival and presentation. Inca server components are hosted at LRZ and are running in a virtualized environment that guarantees efficient load balancing and high fault tolerance.

At the moment Inca is used to monitor the state of selected PRACE user environment software components, including applications, compilers, shells and tools. Inca reporter manager is installed and running on the FZJ JUGENE system. Deployment of a reporter manager on the CEA CURIE system is planned in the near future.

All Inca client and server components deployed in PRACE are at version 2.5. Inca 2.6 release is currently being evaluated in PRACE. Upgrade to the latest Inca release should take place within one or two months, based on the outcomes of the evaluation process.

The Inca configuration is currently being expanded based on the scheme designed during the first year of the project. Inca will be used to monitor the state and functionality of all publicly accessible PRACE components. Individual Inca tests are grouped in the so-called suites configured based on the scope of the PRACE operation sub-tasks. This not only helps to ensure a high quality of service provided but also allows collecting statistics and compiling detailed reports on availability and reliability of PRACE data, compute, AAA, user and other services.

4 Identification, selection and evaluation of new technologies

This chapter describes the work that has been performed within Task 6.3 during the first year of the project. Task 6.3 is focused on: i) performing requirements analysis on the technical and the user level; ii) monitor, assess and select new technologies for Tier-0 to Tier-0 and Tier-0 to Tier-1 integration (i.e. seamless access for users); iii) customise and maintain operational tools and deployed services.

As mentioned in Paragraph **Fehler! Verweisquelle konnte nicht gefunden werden.**, the major part of allocated effort has been dedicated to achieve the following objectives.

a. Set up a formalized process to guide the evaluation of new technologies and the design of new services (see Par. 4.2.1).

b. Collect and analyse requirements from users. Much attention has been devoted in getting users collaborate to the definition of internal technical plans so to ensure deployed services will effectively meet their requirements (see Par. 4.3).

c. Define technical requirements in order to permit PRACE-RI evolve in a consistent manner as new Petascale systems will be integrated into the infrastructure (see Par. 4.3.2).

d. Consolidate existing technologies to continue along the path set out during the preparatory phase to ensure users adopt same interfaces to access PRACE systems.

e. Evaluate new technologies on the base of identified requirements (see Par. 4.5, 4.6, 4.7, 4.8, 4.9, 4.10).

f. Improve the quality level of available services (see Par. 4.4).

In the following paragraphs more details about task activities and achievements of aforementioned objectives are presented. The structure of this chapter is organized as follows.

- Organisation of the task.
- Process description for the selection and deployment of new services.
- Requirements analysis.
- Services certification process.
- Work done on individual service categories (Network, Data, Compute, AAA, User, Monitoring), including a summary table with requirements to address, progress status of ongoing activities and plans for the second year of the project.

4.1 Organisation of task T6.3

Task 6.3 is organized in two different sub-tasks:

- Task 6.3.1 – covers the collection and analysis of requirements on the user and technical level;
- Task 6.3.2 – performs technology watch, assessment, test and deployment and customization of technologies for Tier-0 to Tier-0 and Tier-0 to Tier-1 integration. This

sub-task is divided in seven activities each covering a specific technology area. The table below presents the task's breakdown structure, including leading partners and the contact person for each activity.

Sub-task	Description	Leading Partner	Contact Person	Area
T6.3.1	Requirement analysis	PSNC	Mirosław Kupczyk	Requirement
T6.3.2	Technologies watch	CINECA	Giuseppe Fiameni	
T6.3.2a	Network services	FZJ	Ralph Niederberger	Network
T6.3.2b	Data services	CINECA	Marcello Morgotti	Data
T6.3.2c	Compute services	BSC	Gabriele Carteni	Compute
T6.3.3d	AAA services	SARA	Jules Wolfrat	AAA
T6.3.3e	User services	EPCC	Liz Sim	User
T6.3.3f	Monitoring	LRZ	Ilya Saverchenko	Monitoring
T6.3.3g	Generic	BSC	Gabriele Carteni	Generic

Table 7: Task 6.3 breakdown structure.

4.2 Process description for the selection and deployment of a new service

The process to deploy a new service, which ranges from the collection of users' requirements to the customization of a specific technology to implement missing features, is presented in Figure 9. Process steps are described below.

- **Collect requirements**: concerns the collection and formalization of user's and system administrator's requirements. Based on a closely collaboration with the WP7 for the definition of user's surveys, it is carried out by sub-task 6.3.1. The technical requirements document, presented below in this document, is defined during this step. Collected requirements are filtered and transformed in lower level, more technical objectives before being passed to successive steps.
- **Select technology**: concerns the selection of technologies that might address detected requirements. This selection is based on a formalized process, called ISTP (Internal Specific Targeted Projects) set out in DEISA2 (see Par. 4.2.1). The sub-task 6.3.2 is responsible for this step.
- **Watch new technology**: concerns the watching of emerging technologies that might be of potential interest for the project. The collaboration with other EU projects, such as EMI or IGE, is a key factor for this step to achieve valuable results.
- **Evaluate and Test**: concerns the evaluation and testing of selected technologies in a pre-production environment.
- **Customize**: if the evaluated technology needs to be customized before being offered to users, any customization is performed in this step.

- **Deploy**: concerns with the deployment of selected technologies in production. The deployment of a new service is carried out in collaboration with task 6.2.
- **Certify**: deployed services are certified during this step. The process of certifying services has not been finalized yet and its effective implementation is still under discussion.
- **Maintain**: if any customization has been performed, this step will take care of their maintenance (i.e. bugs fixing, adaptations).

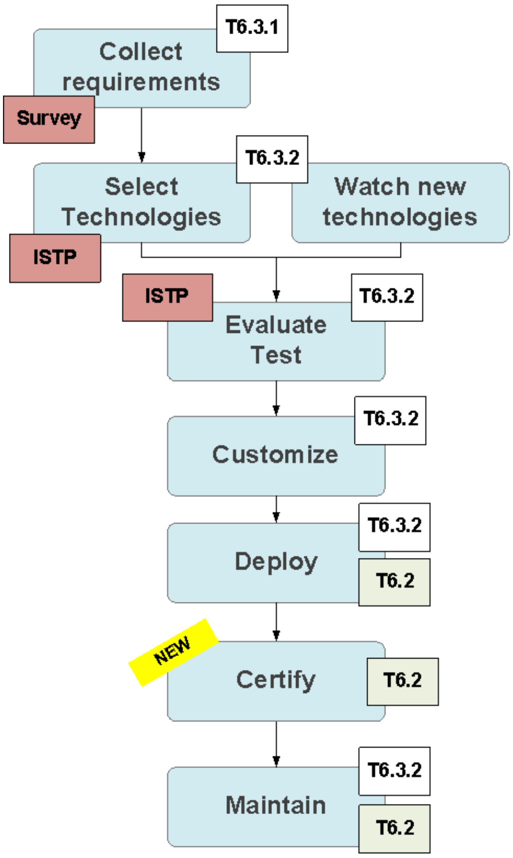


Figure 9: Steps involved in setting up a new service

4.2.1 *ISTP (Internal Specific Targeted Projects) process*

A project-based approach (ISTP) for the evaluation of new technologies was employed with success in DEISA 2. Built upon a rigorous evaluation procedure, it states that every time a new technology, potentially addressing a well-defined set of requirements, needs to be evaluated, a project like approach must be adopted. This should guarantee that all detected requirements are taken in consideration and that clearly and identifiable results are delivered to users in a reasonable amount of time.

The following figure presents the phases that compose the ISTP process. A brief description for each of them also follows.

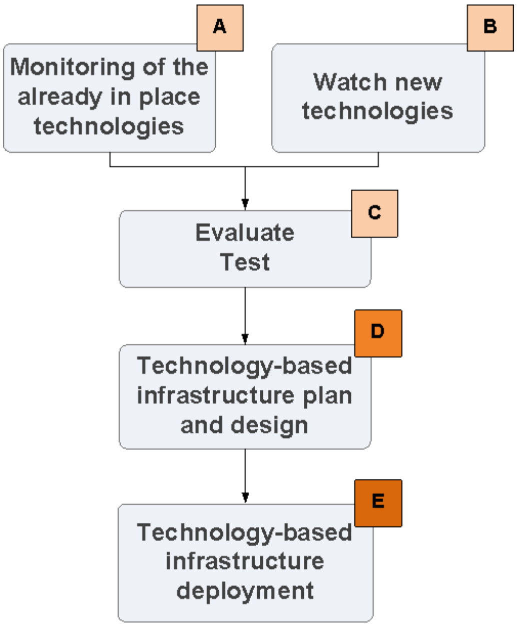

A
Monitoring of the already in place technologies
B
Watch new technologies
C
Evaluate Test
D
Technology-based infrastructure plan and design
E
Technology-based infrastructure deployment

Figure 10: ISTP breakdown structure

A. **Monitoring of the already in place technologies**: monitor the evolution of the technologies already in place.

B. **Watch new technologies**: identify and select new technologies that might either be of potential interest to address specific requirements or enhance the infrastructure.

C. **Evaluate and test**: evaluate technologies for pre-production deployment purposes.

D. **Technology-based infrastructure plan and design**: plan and design specific sub-infrastructure where to deploy the technologies evaluated with success.

E. **Technology-based infrastructure deployment**: deploy and test the above sub-infrastructure on the PRACE production infrastructure.

Constraints of the process:

1. phase A and B can start simultaneously;
2. phase C depends on B. Phase B ends with a GO/NOGO statement;
3. phase D depends on C. Phase C ends with a GO/NOGO statement;
4. phase E depends on D. Phase D ends with a GO/NOGO statement.

The outcomes of each phase are reported in an internal document according to a pre-defined template. The evaluation of each technology is carried on with the collaboration of T6.3.1 (Requirement collection) for ensuring that all requirements have been correctly interpreted, and task 6.2, for ensuring that all the aspects concerning the operation of the service have been also taken in consideration.

4.3 Requirement analysis

Requirements collection and analysis activity is a continuous effort of this task to ensure users access PRACE resources in a proficient and fruitful manner. Taking in account user and technical requirements from day-to-day operations of the system, the PRACE-RI will be enhanced with cutting edge of technology providing a unique persistent pan-European

Research Infrastructure for High Performance Computing (HPC). Steps encompassed during this activity are:

- **gathering of new requirements**: communicating with users, requirements are collected either by direct (i.e. survey) or indirect methods (PRACE User Forum [48], focus group, general discussion, etc.);
- **analysis of collected requirements:** requirements are filtered to remove unclear, incomplete or ambiguous requests and then transformed into lower level, more technical, objectives. During this step, a priority is also assigned to each requirement;
- **recording of requirements:** final requirements, or technical objectives, are provided as input to subtasks work-plan. As not all detected requirements can be addressed, mostly due to effort limitations, only those with the highest priority are considered. Requirements which are not taken in consideration during one cycle, might be reconsidered in the future.

During the reporting period, a reasonable amount of effort was dedicated to collect and analyze user's requirements so to ensure the laying of solid foundations for the evolution process of the research infrastructure.

4.3.1 *User Survey results*

WP6, in collaboration with WP7, conducted a series of surveys of current PRACE partners' HPC systems, the applications running on them, and of current/potential users of the PRACE infrastructure. The first "User Survey" contained questions on the following subjects: System Level, Applications Level, User Level (Application Enabling, Application Usage on the System, User Services) and was carried out between November 2010 and January 2011 targeting all current and potential users communities of all countries that participate in PRACE. Questionnaire results were used by Task 6.3 to indentify technical needs from end users. In order to achieve this goal a two-step process was followed. First step, only questions that were relevant to technology requirements were selected and respective responses made available to subtask leaders. Second step, responses were filtered so that only requests with a desired scalability for their applications being higher than 2048 cores were taken in account. Responses filtering was necessary to identify only users interested in using Tier-0 class resources (154 answers on the original 411 - 37.5%). Consequently, a further analysis was performed on filtered answers to restrict the set of questions that was identified as crucial to draw effective conclusions on required technologies. Final results and further elaboration are maintained in an internal WIKI page.

General observations from above analysis are.

- Tier-0 users (requiring more than 2048 cores) have a good knowledge of HPC systems and perfectly know what are their needs. i.e. they answered more questions in comparison to the full set of respondents.
- Users perceive a great difference between Tier-0 and Tier-1 systems in terms of performance and number of cores (Machine details). On the basis of their opinions, availability of Tier-1 machines is a key enabler for reaching the Tier-0 level, although that difference could hinder their applications to scale up from Tier-1 to Tier-0.
- There is a great heterogeneity in terms of libraries, development tools and applications, which could justify the adoption of something similar to the Common

Production Environment (CPE) of DEISA. At the same time such heterogeneity could require a major effort to maintain such integration layer.

- About 1 user out of 5 could store and transfer 1 TB of data or more per month, so the management of big amount of data deserves attention. In the group of users that require more than 2048 core this percentage increases (see Figure 12).
- Many users are interested in sharing ideas, through WIKI, forum, documents, etc.
- Most part of users would like to have information about their resource usage, if not in real time, at least on daily basis. Available disk space is one of the most important measures (see Figure 11).
- FORTRAN, C and C++, combined with MPI and OpenMP are still the most adopted programming languages and libraries, sometimes in a mixed way.

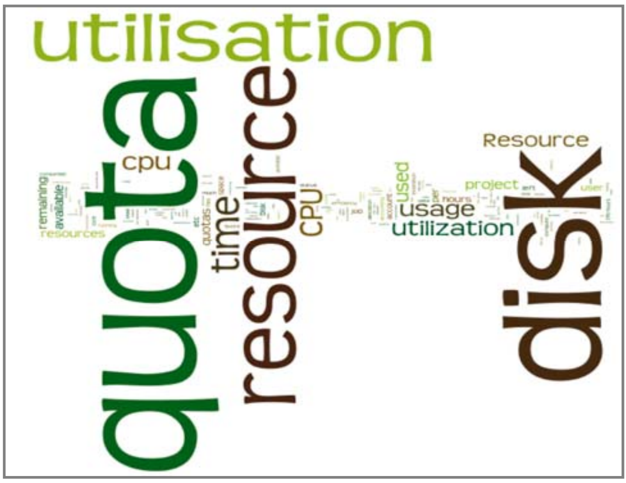


Figure 11: Words cloud of user's requests for resources[3]

- Most popular and well-known batch scheduler are PBS and LoadLeveler. However other batch systems, such as SLURM and SGE, were also mentioned by a good percentage of users (10-20%).
- Only 20% of users have a grid certificate, however in the group of users that require more than 2048 cores this percentage increases to 31%. DEISA users should be familiar with grid certificates, as this PKI is used in the DEISA infrastructure.

Apart from questions that had a predefined set of answers, users were also invited to submit their opinion in free text and all these responses were taken into consideration while drawing final conclusions.

In the following table high-level requirements are presented associated with the list of actions that will be undertaken to address them.

High level requirements	Task plans to address requirements
User Accounting views are important	Accounting system needs to be improved to offer more information on resource utilization status, especially on available disk space.

[3] This picture presents which words had major number of occurrences within free-text responses. "Quota" turned out to be the most mentioned word.

More information on installed libraries	More and better documentation.
Disk space for data and transfer of large amounts of data in and out of the HPC system is important	Improve existing data services for facilitating management, transfer and archiving of scientific data. Evaluation of new technologies is already in task work-plan.
Existence of a variety of architectures to cover different needs of applications but also available expertise	By 2012 four different Tier-0s will be available.
Share of expertise and availability of personnel to support the optimization and even usage of applications	Provide collaborative tools, such as forum. At the moment this is marked as low-priority requirement.
Short waiting time in queues	Tools for reserving resources in advance might reduce waiting time and help users better organize their work. Advance reservation tools were already evaluated during previous projects, DEISA2 and PRACE-PP, and their adoption was considered inappropriate at that time.
Provision of development resources for testing and development before moving to usage of Tier-0 resources	Preparatory projects should help addressing this issue.
Easy authentication/authorization system	X.509 certificate are still perceived as an obstacle. A possible collaboration with EMI project for the development of a unified security token service is envisaged to overcome this issue.
Fast and reliable reviewing/access procedure that can accommodate a variety of projects (i.e. small and large ones).	PRACE AISBL [2] will take care of this request

Table 8: High level requirements and corresponding actions

The outcome of the presented survey was also used as input for the definition of service technical requirements. The following section presents the work that has been done to identify fundamental *Technical Requirements* for production services.

4.3.2 *Technical requirements*

As presented above, one of the goals of Task 6.3 is to collect requirements on user and technical level. Therefore the aim of this section is to present the definition of a preliminary set of technical requirements for the operation part of the PRACE-RI. The methodology exploited here is two-way: the collection of information from previous project documents (i.e. DEISA2, PRACE-PP) on the same specific topic and the gathering of comments from people responsible for the production of PRACE systems.

Definition of a technical requirement

A technical requirement (or non-functional requirement) pertains to technical aspects that a system must fulfil, such as performance-related issues, reliability issues, and availability issues. These types of requirements are often called quality of service (QoS) requirements, service-level requirements or non-functional requirements.

In systems engineering and requirements engineering, a non-functional requirement is a requirement that specifies criteria that can be used to judge the operation of a system, rather than specific behaviours [23]. This should be contrasted with functional requirements that define specific behaviour or functions.

Furthermore, IEEE [24] [26] defines Non-Functional Requirements as "*a software requirement that describes not what the software will do, but how the software will do it, for example, software performance requirements, software external interface requirements, design constraints, and software quality attributes*"

Non-functional requirements characterize the behaviour that is required in functional Requirements which are, instead, gathered in the PRACE Service Catalogue [47]**.**

The PRACE Service Catalogue has been developed in parallel and contains the minimum set of services that PRACE contributors must offer to users. However, as the current version of the document is lacking of an exhaustive description of expected functionality for the services, some improvements will become fundamental in the future to make this work really effective.

Technical requirements activity details

Services that were proposed for the deployment in the production environment of PRACE Tier-0 layer, have been enumerated and described in a PRACE-PP deliverable [42]. PRACE-PP deliverables document the work devoted to services and tools which had been evaluated on the prototypes available at that time. During this project's lifetime, we had an adequate infrastructure to deploy services on Tier-0 machines. The report on the current installation of the common services on Tier-0 machines is presented in chapter 3. During the past year of the project, and as a result of the implementation PRACE AISBL [2], the part of Tier-0 infrastructure is operational and users can apply for computational resources. That is why, the idea of conducting the survey presented above and concerning the Tier-0 users and operating staff about their ordinary work has gathered interest.

Different information sources, such as the PRACE Service Catalogue, PRACE-PP documents, "User Survey" outcomes, User Agreement and internal discussion with operating staff, have been taken in consideration to identify technical requirements. In the following, the most relevant ones for improving PRACE services are presented.

- **Availability**: the degree to which a service is in a specified operable state. Tier-0 service availability is the proportion of time a service is in a functioning condition.
- **Usability:** ISO [26] defines usability as "*The extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use*". Usability is a qualitative attribute that assesses how easy interfaces are to use. Usability cannot be directly measured but must be quantified by means of indirect measures or attributes such as, for example, the number of reported problems with ease-of-use of a system or service.
- **Reliability**: it is defined as: "*the probability that a device will perform its intended function during a specified period of time under stated conditions*". Generally, this is taken to mean operation without failure. However, even if no individual part of the system fails, but the system as a whole does not do what was intended, then it is still charged against the system reliability.

- **Performance**: the overall performance of the system, or any subsystem arbitrary chosen previously. In HPC environment usually one meets the following measures of performance: operations per certain amount of time, number of concurrent processes, the time of execution some part of the work, and so on.
- **Logging**: logging of relevant events encountered on the system. It depends on the operator, what and how detailed data will be stored in the log files depending on the their severity. There is also a need to define, who can retrieve these data.
- **Security**: security includes all steps that are to be taken to secure the system against both voluntary and involuntary corruption. This includes management of users' credentials; encryption of data transfers both internally and across external systems such as the internet; firewalls and protection against any kinds of malicious code attacks including denial of service.
- **Supportability**: specify ease of installation, configuration and testing in terms both of the time to achieve the goal and specific means for achieving it. Consider, for example, installation software and scripts, use cases for configuration and automatic self-testing. Think carefully about how each of these requirements will actually be tested.

According to this list, the definition of core services will be further enhanced adding which technical requirements their implementation will have to satisfy. Where possible, explicit qualitative thresholds (e.g. percentage of service availability) will be set. This information will be collected in a reference document called "*PRACE Services Technical requirement*" and shared with partners.

In addition to the list presented above, which mainly focus on general, long-term criteria, a further list of basic requirements have been also identified to give a pragmatic view on changes that will affect current services.

- Service, or product, must provide a document with a description of the functionality it supports (this information is partially contained in the PRACE Service Catalogue).
- Service, or product, must provide an administrator guide describing installation, configuration and management of the service.
- Service, or product, must have a mechanisms for starting, stopping and querying the status of the service, following the hosting OS init scripts conventions.
- Service, or product, must provide monitoring probes that can be executed automatically by the monitoring system (INCA).
- Service, or product, must provide ways or recording the use of resources within the infrastructure for the PRACE account capability.
- Service, or product, must maintain a good performance and reliability over long periods of time with normal load.
- Service, or product, must not create world-writable files or directories for security reasons.
- Service, or product, must be well-supported by developers group or company.
- Service, or product, should possibly be open source and based on widely adopted standards to facilitate the interoperability among different infrastructures.

Of course, most of the requirements listed above already influence the evaluation of new technology and their applicability is detected during ISTP steps.

It is worth mentioning that the main goal of this activity is not to make technical requirements immediately mandatory but rather introduce quality concepts in the management of the PRACE infrastructure as a continuous process consistent with the contents of the Contributors Agreement.

To be as effective as possible, for each core service (see PRACE Service Catalogue), a table with the name of the technology implementing its functionality and a preliminary list of non-functional requirements has been defined (see Appendix A).

During the first year of PRACE-1IP project, several technologies have been deployed on Tier-0 and run into the production mode. Some of them have been tested using different configurations of middleware services – the outcome of that evaluation is included in the corresponding tables. The tests were being performed both on the PRACE-RI and DEISA infrastructures. In some cases, DEISA infrastructure has been employed as the PRACE test-bed. DEISA offered the computational environment with similar architectures and services but in small configuration, as some machines belong to the PRACE Tier-1. For the sake of this work, we recap the current achievement in tables presented in Appendix A. The meaning of the corresponding table fields is the following.

- **Service Name**: the general name of the service, widely recognizable, usually self-explaining. By these names one can refer to particular service in other PRACE documents, e.g. Service Catalogue.
- **Technology**: the specific tool, or solution, implementing the Service functionality of different types. This tool has been tested against the fulfilling the requirements.
- **Functional requirements**: the main requirement(s) which the service is devoted to.
- **Non-functional requirements:** the set of requirements derived from the testing of particular **technology** and aforementioned sources.
- **Evaluation**: the short report on the testing of Technology, or other alternative solutions. We do not go through existing technology to re-evaluate them, rather than to identify attributes of the non-functional requirements. The measurable attributes will be defined in the 2nd year of the project and after thorough discussions the reasonable thresholds for these requirement will be set (e.g. % of availability, reliability).
- **Resource Requirements**: the HW and SW requirements derived from **Evaluation,** fulfilling **the non-functional requirements,** which impose the optimal characteristics of the particular **Service**. This field has been intentionally left empty – it will provide some hints regarding the adequate configuring services, but it has not been our intention to provide the manual on configuration which is assumed to be provided by the technology provider.

Currently adopted technologies will not be re-evaluated but rather production services will be measured so to evaluate their performance and thus set a threshold for them (i.e. % of availability, reliability).

During the second year of the project, technical requirements will be made progressively mandatory and, if needed, revisited according to site's capabilities, performance of current deployed services, available effort. The monitoring system will be extended to measure the availability/reliability level of current services in order to set reasonable thresholds for involved technical requirements.

4.4 Services certification process

Service certification is a function to ensure that any deployed service is compliant with PRACE policy and offers a good level of quality supported functionalities. A successful certification should be mandatory for any service to be accepted into the PRACE-RI. This would enable users to access reliable, well-documented, easily manageable services, regardless of hosting site. Any service should be certified before entering the production level as its configuration could not be correct, not support all detected functionalities, or not provide satisfying performance. In general, the definition of a certification process would be fundamental to understand whether PRACE's offering meets user's expectations.

To achieve this goal, a certification process proposal has been prepared building upon the following:

1. ensure that services are fully documented: services should be supported by user and system administrator manuals;
2. ensure that services are correct: expected functionalities are correctly supported. The list of functionalities (i.e. functional requirements) expected by each service will be part of the PRACE Services Catalogue;
3. ensure that services are robust and reliable: technical requirements (e.g. non-functional requirements) are sufficiently satisfied. The list of technical requirements, non-functional requirements, which are expected to be supported by each service will be included in the technical requirements document (par. 4.3.2)**,** still under discussion;
4. ensure that services are offered at a certain level of quality checking that quality standards are satisfied. The list of quality standards which are expected to be satisfied by services are currently not fully defined in any document, although it is highly probable that they will be included in the User Agreement.

Any control is foreseen to be implemented through a check-list: once a check in the list is executed with success, the corresponding item is checked out from the list. Checks could be defined by people from task 6.3 while their execution put in the hands of task 6.2. This separation would also guarantee a clear distinction between who defines quality levels, namely the Quality Assurance, and who performs the checks, namely the Quality Control.

Due to the possible overhead generated by service certifications, they will be scheduled only during critical phases of services life-cycle:

- before a new service enters the production level;
- after any major change to the configuration of the hosting site.

The certification process will not replace the monitoring system, but rather the integration of the two systems is fundamental to reach valuable results. The major differences between them are:

- **certification**
  - its aim is to control that deployed services comply with requirements and meet quality standards;
  - it is performed only during critical phases of service life-cycle;
  - as its checks intent to simulate user's interaction, it might consume precious computational resources;
- **monitoring**

- its aim is:
    - to control that production services are up and running;
    - to support the operation of the infrastructure;
    - to collect statistics on services lifetime and utilization (i.e. uptime of the service, mean time between failure, number of access, etc.);
    - to control that technical requirements are meet (i.e. availability, reliability, etc.)

To show how the service certification process could effectively be implemented, in the table below a set of certification tests are presented for the „*Unified access to HPC Infrastructure*" service (see PRACE Service Catalogue [47]).

Service name	*Unified access to HPC Infrastructure*
Requirement	**Certification tests**
• Provide users access to PRACE computational resources • Authentication protocol should be based on an open standard	Check if the service permits the authentication through X.509 certificates of three different users, belonging to three different CAs.
• High availability/reliability: the service must be available 90% of the time	Service must not show performance degradation after 2 days of operation. Parameters to check are: • stable memory and CPU usage; • response time should remain stable during the period of activity (they should be as good or better than at the beginning of the test for similar requests).
• Authentication should be two-way: Client-to-Service, Service-to-Client • No credentials passing in clear • Full credentials not stored for long period. Stored credentials should have an expiration date • Opportunity to configure with credentials expiration, history, and intruder lockout	• In turn remove the server and client public key and check that authentication step fails • Check the existence of world writeable or ownerless files in the system • Mark an account as "expired" and check if it can still access the system
• Logging: date, time, source IP, username for any tentative access should be recorded	• After having accessed a machine, check if the required information is logged on the remote machine

Table 9: Sample tests to certify the "Unified access to HPC Infrastructure" service.

4.5 Network services

Within PRACE, it was planned to bring further components of perfSONAR (esp. BWCTL measurement service) as new technology into production. The evaluation revealed that the software is not ready for production yet. A new version which should fix those problems will be released in August/September 2011 by GEANT3 project [27]. Therefore, an evaluation of the new release will be done in future by Task 6.3.

Network services summary table

Requirement	Candidate technologies/solutions	Progress Status	2^{nd} year plan
Consolidate existing technology (perfSONAR suite)	BWCTL measurement service (it is part of the perfSONAR suite).	Considered but to be evaluated when the new version of the software will be released.	Evaluate the BWCTL measurement service.
Provide aggregate Network status information (see PRACE Infrastructure Status)	Not identified yet. Currently there are three candidate technologies: INCA, DMOS and perfSONAR Visualization portal.	No technologies have been evaluated yet. This activity is currently working on defining the requirements for the service (PRACE Infrastructure Status). It is carried on in collaboration with the monitoring and accounting sub-task.	Select the right technology and implement defined requirements.

4.6 Data services

Being able to manage huge data sets produced by simulations running on Petaflop machines is one of the main challenges of PRACE. From the user survey, it has emerged that 1 to 5 users produces at least 1Tb of data per month.

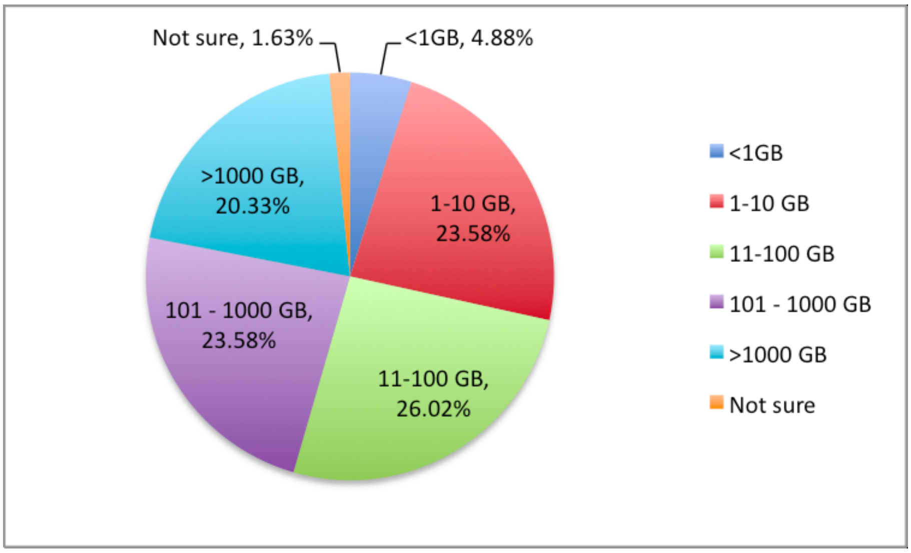


Figure 12: Minimum amount of disk space required for production jobs (taken from user's survey results)

Within the first year of project, technologies put in production during the preparatory phase of the project have been considered still valid to address user's demand, however, as new requirements arose, the research for new technologies continued on the following areas of interest.

- **Reliability of data transfers:** it has emerged from many use-cases and also from survey outcomes that often users face many problem during their tentative to move data to/from the machine where they are performing their computations. This happens because there is no way to guarantee that their transfer are going as expected nor there is not any agreement on that (i.e. some sites have CPU-TIME limits enforced on login nodes which will interrupt any transfer exceeding the limits). Two technologies which can enable the opportunity to schedule transfer and so overcome mentioned limits have been already selected and their evaluation will occur during the second year of the project:
  - **UNICORE [20]:** it is now able to schedule file transfers within the latest server release 6.4.0, currently available for downloads. The UNICORE command line client (ucc 6.4.0) already supports this functionality; the graphical client will be able soon, as well. A test-bed installation will be provided by FZJ to permit other partners this new feature;
  - **Globus Online [28]:** it is a cloud service, hosted by Amazon cloud infrastructure. Its main aim is to provide a web-based interface to GridFTP. It is made of two major components: the web service and a "client" (Globus Connect client) which acts like a GridFTP server on the local machine in which it is installed. The client is, nowadays, GUI only, but a CLI version will soon be available which will be usable for transfer scheduling. A preliminary evaluation of this technology has reported some security concerns that, with the collaboration of the PRACE Security team, will be further investigated in order to understand the impact they could have on the PRACE infrastructure.
- **Data Archiving**: providing the necessary middleware for a Data Archiving facilities available for each Tier-0 sites is the real new challenge of PRACE-1IP in this field. Users accessing Tier-0 systems and generating tens of Terabytes of data must be provided with the opportunity to archive their data and retrieve them even after the end of the project in an efficient and easy way. Two data-archiving technologies have been already selected and their evaluation will start in the future:
  - **dCache [49]**: it is a disk-pool management system with a SRM interface [51], jointly developed by *DESY* and *Fermilab*. It offers an intuitive mechanism for storing and retrieving huge amounts of data, distributed among a large number of heterogeneous server nodes, under a single virtual file-system tree with a variety of standard access methods. Depending on the persistency model, dCache provides methods for exchanging data with backend (tertiary) Storage Systems as well as space management, pool attraction, dataset replication, hot spot determination and recovery from disk or node failures;
  - **OpenStack Object Storage [50]:** it is an open source software for creating redundant, scalable object storage using clusters of standardized servers to store petabytes of accessible data. Its main goal is to offer a long-term storage system for a more permanent type of static data that can be retrieved, leveraged, and then updated if necessary.
- **List of available services related to the operation that the user wants to perform**: Users are not aware of all available services, how to get the best from them and in which situations should they be used. To help them select which option best fits their

needs, the available documentation will be improved to present, for the most frequent data-transfer scenarios, which combination of services and configuration parameters, is the best choice for them to adopt.

- **Quota information**: Due to the heterogeneity of available file-systems (i.e. GPFS, AFS, Lustre, pNFS etc) in use, the quota enforcement might be different from site to site and from one system to another. This forces users to learn new way to retrieve their quota information every time they access a different system. It would be therefore interesting to have an abstraction layer to wrap all these information so that a single command can provide the necessary information careless of the underlying file-system type and configuration. The possibility to develop this abstraction layer will be evaluated in collaboration with the AAA sub-task and, if possible, integrated as part of the accounting system.

Data services summary table

Requirement	Candidate technologies/solutions	Progress Status	2nd year plan
Improve data transfer service reliability	Improve documentation in order to help users select which service, among those available, best addresses their needs.	A comparison table containing available services and possible configurations has been already prepared.	Complete the comparison table and integrate it into the official documentation.
	New UNICORE 6.4 features	To be evaluated, waiting for a test-bed installation to be set-up.	Complete the evaluation.
	Globus On-Line	Under evaluation. Security concerns arose during the evaluation will be submitted to the PRACE Security Team.	Complete the evaluation.
Offer a data archiving solution	dCache	Considered, to be evaluated.	Investigate the possibility to carry on the evaluation according to available effort.
	OpenStack	Considered, to be evaluated.	Investigate the possibility to carry on the evaluation according to available effort

4.7 Compute services

The study for providing a common interaction layer to all computing facilities available in PRACE-RI is the main focus of this activity.

Common issues, which more than others affect this analysis, are:

- different software for resources management and usually tailored to a specific facility/platform;
- customizations in place for improving performance and/or user interaction;

The adopted strategy is to move to a higher layer and find a middleware solution that might act as intermediate among users and systems for a common job management interface. This is in line with the outcomes of the preparatory phase of PRACE [29] [30].

The work-plan on this service category reflects a first set of user requirements collected in this first year and described in the next section.

4.7.1 *User survey for compute services*

Three questions related to compute services were included in the User Survey submitted by Task6.3:

1. which of the following batch-schedulers are you familiar with?
2. which of the following grid middleware are you familiar with?
3. does your application require a workflow system to be executed? Which of the following systems are you familiar with?

Relevant information out of 123 replies is:

1. only the 10% of users is familiar with SLURM, which is the resource manager used by CURIE system available at CEA;
2. at least the 20% of users is familiar with UNICORE while half of them don't have any expertise with grid technologies. From these two figures we can extrapolate that at least the 40% of grid aware users are familiar with UNICORE.
3. almost all users are not interested to solutions for managing workflow of jobs.

It is important to say that users don't belong to a specific target; they come from different scientific fields and use different technologies for using computing resources.

Anyway these figures represent a good starting point. Surveys submitted to specific target of users are needed to complete the profile of a typical user interested in compute services provided by HPC systems.

4.7.2 *UNICORE enhancements*

UNICORE version 6.4.0 has been released on April 2011. At the time of this writing, only a plan of new features has been defined. In particular the possibility for a user to select a specific queue of a remote Batch Scheduler System and the scheduling for a certain time of a job submission. A new survey targeting UNICORE users is under evaluation to be submitted in the upcoming UNICORE Summit which will take place will take place at Nicolaus

Copernicus University, Torun, Poland, on July 7th - 8th. Apart of these new features, the evaluation of the workflow engine, taking into account it has already evaluated in DEISA, has been planned.

4.7.3 *Evaluation of alternative solutions*

The ProActive Parallel Suite [32] is the first solution for which a preliminary contact was taken by responding to an explicit contact from vendors.
A first demo has been executed with the software developers. ProActive is Java grid middleware for parallel and distributed computing. It is developed by the OW2 Consortium [33] and released as open-source software under the GPL license.
ProActive provides a comprehensive framework and parallel programming model to simplify the programming and execution of parallel applications running on distributed resources.
Part of this suite is a Resource Manager, which allows at managing in real-time resources on the Grid and user activities. It leverage on XML language to create an abstraction layer with remote Batch Scheduler System. Security is delegated to the SSH protocol.
Further contacts have been planned with the representatives of ProActive to follow an in-depth analysis, in particular on security, interoperability and installation.

Compute services summary table

Requirement	Candidate technologies/solutions	Progress Status	2nd year plan
Offer a workflow engine for job submission	UNICORE Workflow system	Considered and already evaluated in DEISA.	Collect more requirements from UNICORE users during the upcoming UNICORE Summit before moving the technology in production.
Provide a common interaction layer to computing facilities	ProActive	Partially evaluated but considered useless to meet current user's requirements.	No plan.
	UNICORE 6.4	Considered, new features to be evaluated.	Evaluate new UNICORE features.
	Local Batch Scheduling Systems	No evaluations are needed for theses technology as they are selected by hosting partners.	Improve documentation according to new batch scheduler system available in PRACE.

4.8 AAA services

4.8.1 *Enhancement of accounting facilities*

Several enhancements for the accounting facilities were developed and evaluated as part of the new technology watch in the DEISA2 project. Some of the results were taken in production, e.g. the publishing of summary records by the Apache/CGI interface at sites. Other results became available in the last project year and were proposed as a production service for the PRACE-RI. Two enhancements will be evaluated by this task:

1. the provision of a central accounting database, based on the Grid-SAFE accounting framework developed by EPCC [15];
2. the provision of budget information to users.

4.8.2 *Central accounting repository*

In ***Figure 13*** the addition of the Grid-SAFE based enhancement to the existing accounting facilities is shown in the grey area. For details on the production facilities see section 3.5.4.

The basic set-up of the new facility is that summary records are periodically sent to a central database. Currently monthly summaries are stored, but the period can be changed. A web interface exists which can display different reports, based on the permissions granted to the requestor and using X.509 certificates for the authentication and authorization.

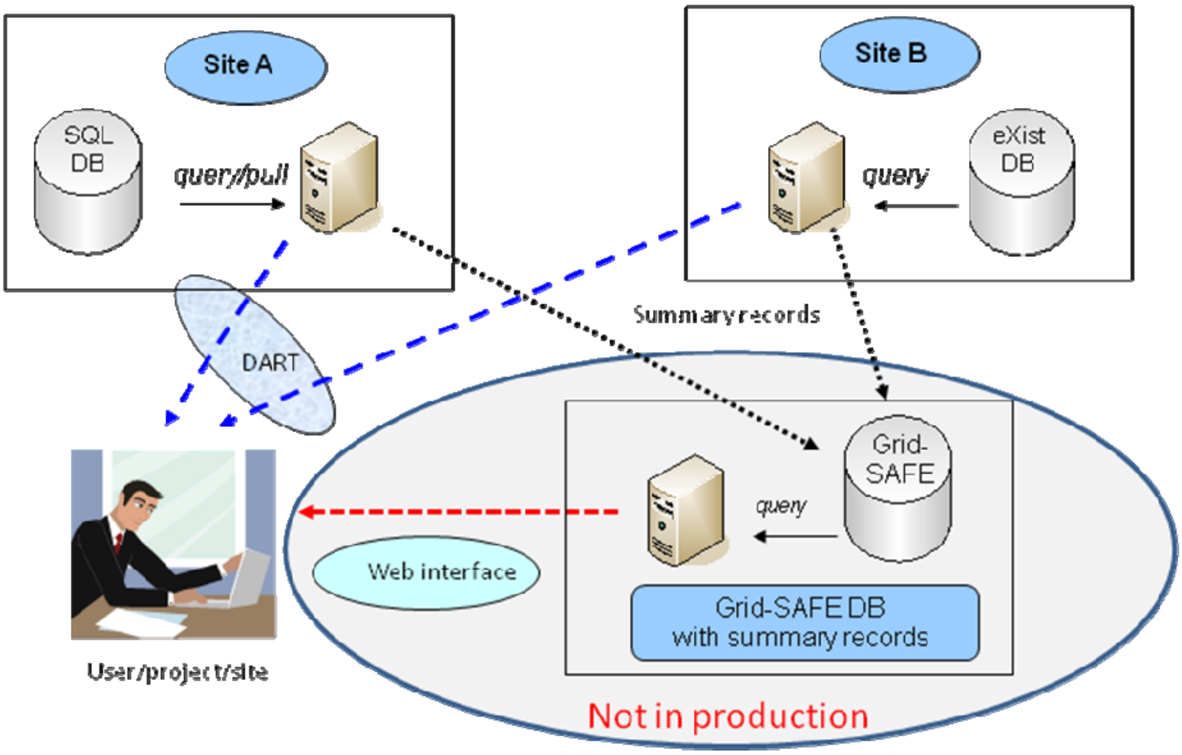


Figure 13: Accounting architecture with proposed Grid-SAFE facility

For the current month at least once a day the summary records are updated, so the central database will always have up-to-date information on the most recent usage too.

There are mainly two reasons to provide these new facilities in addition to the existing facilities:

1. the centralized solution is less resource consuming because the query of each of the distributed accounting databases for summary records (via the CGI interfaces) is triggered only once after specific intervals by the hosting site, usually when new user records are available. All the output data which is obtained via the Grid-SAFE web

interface is produced at the server side. So, the load on the site local servers is less. Also the storage requirements are less, because only summary records are stored;

2. the web interface is easier to use compared to using the DART client, the user does not have to manage a locally installed client. Basically both interfaces, DART and the central web interface, provide the same information. The only difference is that DART can provide more fine-grained information for short periods.

It is agreed to prepare the central accounting repository for production. EPCC is prepared to operate the server that will host the database. EPCC already operated the test server which was used to evaluate the facilities. For the web interface some additional effort is needed to improve the functionality and the manpower for this must be identified.

Because personal data is involved in accounting records the consequences of storing these on a server outside the domain of the site, which provides the data, must be clarified. Users have to agree that for accounting purposes personal data can be provided to entities, which need this information by signing the AUP (Acceptable Usage Policy) as part of the user agreement. However, the legal consequences for the sites must also be investigated. The EU directive 95/46/EC [16] must be followed and national legislation must also be considered.

4.8.3 *Budget information*

Users and PIs are interested not only in their past usage but also in the remaining budgets. The allocated budgets can already be published in the user administration system on a project level per system. In the LDAP schema attributes are defined for the resources which are allocated to a project.

A version of DART was developed which displays the available budget after the usage is subtracted, using the budget information from LDAP and the usage of the queried period. This does not take into account the usage outside the queried period. For example, if project A has an allocation for six months and the user only asks for the usage of months two to four, then the usage for the first month is not included in the budget calculation. A new functionality would be more appropriate, where a user can query the budget status and where always the whole period of the allocation is considered. A complication is that sometimes the allocation is split in intervals of the total allocation period, and this also should be taken into account. It will be investigated if a version of DART with budget selection features can be further developed for evaluation.

Also the web interface to the central repository can be adapted to display the budget information. This will be planned as part of the further development of the web interface.

4.8.4 *Improvement of AA facilities*

The authentication and authorization facilities fulfil the requirements of the infrastructure. The management of X.509 certificates is not always easy for the end user: 1) client tools have different requirements for the format in which the certificate is used and users have to translate the certificates; 2) browsers have different ways in which the certificates are managed.

All infrastructures relying on X.509 certificates have to cope with these limitations, so it is important to evaluate what is done in other infrastructures and also what middleware initiatives there are in the area of AA in projects like EMI [17] and IGE [18].

Larger collaborations of scientists will not be using just one infrastructure but will have access to resources on different infrastructure. For these communities it is of interest to be able to use these facilities in a transparent way. For authentication and authorization this means that the security tokens used are accepted and trusted by all. In the first place the technologies used must be able to use the tokens, but secondly the information provided by the tokens must be trusted to come from a reliable source. The latter requirement is addressed by the PRACE security forum and is not further discussed here.

Security tokens can be for instance X.509 certificates, proxy certificates or SAML assertions. It depends on the service what kind of tokens can be accepted and so it may be needed that a token is translated between the different representations if the same token is to be used for different services. Several initiatives exist to develop and to evaluate such functionality, which is referred to as a Security Token Service (STS).

This task in the first place must identify the use cases for the STS functionality, primarily based on user requirements for the collaboration with other infrastructures. Feedback to the EMI STS middleware development activity is given that the use of LDAP attributes as input for security tokens is of interest for our infrastructure. In general this subtask must keep in touch with the developments in other infrastructures and projects to identify the enhancements which can be useful for our infrastructure.

4.8.5 *Evaluation of the user administration service*

The LDAP facilities in use for the administration of user and project information are the result of many years of experience of the DEISA project. Started in 2004, the facilities have seen the addition of many new attributes to enhance the functionality. Examples are the attribute for the systems on which projects have been allocated resources and the attribute for the access rights to accounting information. There are two reasons to evaluate the current design of the namespace used:

1. the user information is attached to an account, which means that user specific information like contact details has to be replicated for each account that a user has.
2. LDAP today is used by many partners also for the management of the accounts in their local environment and the PRACE LDAP information is replicated to the local environment. The replication will be easy if the namespaces of the different environment match as much as possible, so no complicated copy actions have to be used. Also the use of standard schema for the research communities would be interesting, see for instance the SCHAC schema [19].

The requirements for the common infrastructure will be reconsidered and the experience of local environments will be used too. The requirements for policies for attributes can also influence the design and must be defined. For instance information about a user must be destroyed in some countries after some time that the accounts of the user expired. At the same time information about accounts may be needed for a longer period.

AAA services summary table

Requirement	Candidate technologies/solutions	Progress Status	2nd year plan
Enhance the accounting system to facilitate user access budget information (see	GridSAFE	Evaluated in DEISA but not moved in production yet. New enhancements have been designed and	• Deploy in production the current release of GridSAFE portal as it already satisfies a sub set of detected

also PRACE Infrastructure Status activity)		are ready to be implemented.	requirements. • Implement designed enhancements.
			Investigate possible solutions to implement an abstraction layer over existing file-system technologies to provide aggregated disk quota information (see also par. 4.6).
Improve the authentication and authorization facilities	Security Token Services (STS)	• Collaboration with EMI project established. • No technology solutions are available at this point in time.	Identify use cases for STS functionality and provide requirements to EMI project.
Enhance the current user administration service (based on LDAP)	SCHAC Schema	Considered but to be evaluated	Evaluate the adoption of SCHAC schema for the LDAP

4.9 User services

4.9.1 *PRACE Common Production Environment*

For the current software installations, Modules has a huge weakness which is that it considers each software package as an independent one and does not offer a possibility to manage the dependencies between them.

Unfortunately, on most platforms this really matters, as soon as there are for instance several flavours of Fortran or C compilers to maintain together, or different MPI implementations, which require to use dedicated versions of some other libraries, etc. This drawback led various people to try to add a possibility to manage the dependencies between software. Various solutions have been implemented and DEISA uses the withdrawn feature of the Tcl version of Modules to manage some of them (but not all).

A key limitation at present is the conflict between the PRACE Modules Environment and the modules environment deployed by Cray as standard on all of their systems. This causes some issues for the upcoming Tier-0 system Hermit at HLRS. Cray chose since many years to use the Modules tool to provide the access to the different versions of all their user software, but in a special way to manage the dependencies between software, as previously explained, which is fully incompatible with all the other solutions developed to circumvent this problem, including the DEISA one. Practically the DME cannot be used today on the Cray systems.

In addition to this, Cray has added some features to the Modules tool that it uses (based on the C version) and now delivers its own implementation.

A small team of experts will be formed to investigate the options available for deploying a PRACE Modules Environment (PME) on Cray systems. This may include customizations to the existing modules environment. Alternatively the SoftEnv [38] environment as used by TeraGrid may be used as an alternative to the current PME. The modules environment used by TeraGrid will also be examined to understand if this is a viable option for PRACE also. This activity has not started yet due to the unavailability of the Hermit system.

4.9.2 *PRACE Helpdesk*

With the exception of the investigation into the creation of a secondary failover TTS system, no additional services or technology changes are planned at this time.

4.9.3 *Visualisation Services*

At this point in time no further visualisation services are planned. The usage of the service available at LRZ and feedback from users will be used to ascertain if any additional services are required.

User services summary table

Requirement	Candidate technologies/solutions	Progress Status	2nd year plan
Enhance the PRACE Modules Environment to integrate CRAY system.	SoftEnv	Considered but not evaluated yet.	Evaluate the technology and investigate its adoption in PRACE.
	Modules system	Currently in production.	According to the outcomes of the SoftEnv evaluation, further investigate the possibility to adapt the current Modules system to CRAY specifics.

4.10 Monitoring services

Monitoring solutions deployed in PRACE collect detailed information about the state of the e-Infrastructure and are used by the operation team to ensure the high quality of service provided. PRACE users could also benefit from access to the monitoring information.

However, at this moment, data collected by both network and user-level monitoring tools is only available to the PRACE staff members. The prime focus of this sub-task is to provide a graphical interface for displaying monitoring information to the PRACE users.

Data presentation capabilities available in Inca [14] and perfSONAR applications cannot be used by the PRACE users as is for several reasons. First, users are normally interested in a subset of the available monitoring data.

For instance, a PRACE user would only want to view monitoring information describing compute resources he or she has access to. Secondly, a part of the collected monitoring data should not be shared outside of the PRACE project due to security concerns. None of the currently deployed tools implement the desired functionality. As such, one of the main goals of this sub-task is to discover and evaluate existing solutions that are able to satisfy the requirements described above.

Another area of interest is management of resource and service maintenance information. At this moment PRACE does not operate a tool that can be used for announcement and documentation of resource and service maintenances. During the course of the DEISA project DMOS (DEISA Maintenance Information Organisation System), a tool for management of maintenance information, was developed. The tool was meant to replace a Wiki-based solution that lacked the necessary functionality. DMOS will be further evaluated within this sub-task and, if necessary, customized for use in PRACE.

One of the most important requirement that emerged from the user survey, also recognized during task internal discussions, is the possibility for users to access information concerning the status of the PRACE infrastructure, including the availability of provided services.

Scheduled maintenance of the systems, failures, network problems, resource overloading are all valuable information that could help users better organize their work and allow them understand what are the reasons behind the problems they might have accessing the infrastructure. The aim of the "**PRACE Infrastructure Status**" activity aims to define which information could be shared with users (some information must be kept confidential), which sources provide them and how to integrate existing systems (i.e. monitoring, network reporting, accounting system, etc.) to form a unified point-of-access for the users.

This activity, which has not started yet, will run across almost all sub-tasks to draw on the synergy of their work.

Monitoring services summary table

Requirement	**Candidate technologies/solutions**	**Progress Status**	**2nd year plan**
Enhance the monitoring system in order to facilitate users access infrastructure status information (PRACE Infrastructure Status)	Not identified yet (could be INCA or DMOS)	No technologies have been evaluated yet. This activity is currently working on defining service requirements. It is carried on in collaboration with the network and accounting services sub-task.	Select the technology to adopt and implement defined requirements.
Provide a tool to track systems maintenance information	DMOS	Already evaluated in DEISA but not in production.	Further evaluate the DMOS technology and, if necessary, customize it to meet PRACE needs.
Extend the monitoring system to collect information on production	INCA	Not started yet.	Evaluate required extensions and proceed with their implementation.

service performance			

4.11 Summary

The following table gives an overview of major activities which are ongoing in task 6.3 and that will start during next months, including their priority (Low, Medium, High).

Each activity is lead by a specific partner and carried on in collaboration with few other contributors. Activity deadlines are expected to be set by the end of June 2011.

General services				
N.	Description	Leader	Task Priority	Contributors
1	Go ahead with the services certification proposal	CINECA	High	ALL
2	PRACE Infrastructure status: • define which type of information on PRACE infrastructure status could be shared with users; • define how to make such information available (e.g. web-site, command-line, Grid-SAFE portal, etc.); • integrate the monitoring, the accounting and the reporting systems together.	CINECA	High	LRZ, FZJ, SARA, *other*
Requirement analysis T6.3.1				
N.	Description	Leader	Task Priority	Contributors
1	Take forward the technical requirements document, elaborate its contents also taking in consideration the services certification concept.	PSNC	High	ALL
Network services T6.3.2a				
N.	Description	Leader	Task Priority	Contributors
1	Adoption of perfSONAR solution in PRACE: • installation of network monitoring scripts (iperf, BWCTL, etc) on CURIE system; • integration of perfSONAR with the monitoring system(see PRACE Infrastructure Status activity)	FZJ	High	CEA
2	Deployment of the perfSONAR visualization portal	FZJ	Low	
Data Services T6.3.2b				
N.	Description	Leader	Task Priority	Contributors
1	Improvement of data transfer reliability/adoption: • documentation enhancement; • users should know how long a data transfer for a given file size could take; • users should know what the file size limit for a give connection bandwidth could be.	CINECA	High	EPCC, CSC
2	Data transfer scheduling facility: • evaluate the new UNICORE features for	CINECA	Medium	KTH

	scheduling data transfers; • evaluate the Globus Online [28] service.			
3	Data archiving facility: • evaluate the dCache software; • evaluate the OpenStack software.	CINECA	Medium	KTH, HLRS
Compute Services T6.3.2c				
N.	Description Plan	Leader	Task Priority	Contributors
1	Conclusions of ProActive solution evaluation (check whether there are PRACE users requiring it).	BSC	Medium	FZJ, CINECA
2	BSS Inventory: • improve the documentation concerning the utilization of batch scheduling systems available on PRACE systems.	BSC	Medium	CEA
3	Preparations of a survey for the UNICORE Summit to gather more requirements from users.	BSC	High	FZJ
4	Evaluation of new UNICORE features provided by 6.4.0 release.	BSC	High	FZJ
AAA Services T6.3.2d				
N.	Description Plan	Leader	Task Priority	Contributors
1	Enhancement of the Accounting system: • deployment and maintenance of the current release of the GridSAFE portal; • development of missing features in GridSAFE; • development of missing features in DART; • tune LDAP configuration.	SARA	High	PSNC, GRNET, CINES
2	Enhancement of the Authentication/Authorization system: • submit requirements for the Security Token Service to be submitted to EMI.	SARA	Low	IDRIS, CINECA
3	Enhancement of the user administration system: • evaluation of SCHAC schema for LDAP	SARA	Low	
User Services T6.3.2e				
N.	Description	Leader	Task Priority	Contributors
1	Investigate on how to adapt the PRACE Common production environment to integrate the new HLRS Cray machine: • evaluate the SoftEnv system; • evaluate alternative solutions, mainly customizations made to Modules system.	EPCC	High	IDRIS, HLRS
Monitoring Services T6.3.2f				
N.	Description	Leader	Task Priority	Contributors
1	PRACE Status: • investigate on how to make monitoring information accessible to users in an aggregate way.	LRZ	High	FZJ
2	Offer a solution for management of resource	LRZ	Medium	

	and service maintenance information: • adapt the DMOS system to PRACE needs; • deploy the DMOS system.			
3	Extend the actual monitoring system to collect information on production service performance (i.e. reliability/availability)	LRZ	Medium	
4	Evaluate alternative monitoring technologies	LRZ	Low	

5 Internal Services

Setup and maintenance of websites, the operation of databases, system status monitoring, trouble ticket system, source code repositories and wiki are all examples of generic/internal services.

In this first year an initial classification of these services has been made. Some services have been organised into specific service categories, this is the case for the Trouble Ticket System and the Authoring Environment for easily producing and publishing user documentation.

Following table summarises all internal services that are currently considered by WP6.

Service	Category	Responsible	Status	Note
WebSite Maintenance	Generic/Internal	PRACE AISBL	Available	PRACE-RI website is managed by CINES
System Monitoring	Generic/Internal	WP6	Available	Wiki page is used for Software deployment monitoring. Application Monitoring is part of Monitoring Services. Network Monitoring is part of Network Services.
Trouble Ticket System	User Services	WP6 (CINECA)	Available for Nov/2011	Includes first and second level of user support.
Source Code Repository	Generic/Internal	WP6 (SARA)	Available	Used by WP7 for software development and also as database for benchmark results. Authentication with X.509-based certificates.
Produce and Publish Authoring Environment	User Services	WP6	Not Available	Work in progress. Main issue to be faced is the integration with the framework used by the website. Incompatibility issues are in place.
Web-Based Collaborative Environment (WIKI)	Generic/Internal	WP6 (FZJ)	Available	Authentication with X.509-based certificates.
Document Management (BSCW)	Generic Internal	WP6 (FZJ)	Available	Authentication with username/password-based credentials.

Table 10: List and classification of Internal Services

5.1 Collaborative services

Online collaborative services are obviously essential for the success of every project involving persons working at different sites/locations.

In this first year, two collaborative tools have been provided and used for daily operations: a common workspace for document management (BSCW) and a wiki-based website implementing a collaborative environment (WIKI).

Both of them are documented in the PRACE WIKI [3].

PRACE WIKI is used for sharing information and documentation, activity tracking and coordination and for implementing tools like the service deployment monitoring on Tier-0.

BSCW is used to manage internal documents such as deliverables, internal reports and minutes of videoconferences.

5.2 Technical services

The only technical service currently provided is the Source Code Repository service, hosted by SARA. It is implemented by using Subversion [46], a centralized version control system characterized by its reliability, the simplicity of its model and usage and its ability to support the needs of a wide variety of users and projects. Authentication requires a trusted X.509 certificate.

It is mainly used by WP7, which focuses on software development.

Until now, SVN is also used as a repository for benchmark activities. This is an accepted workaround to allow WP7 to store results of benchmarks in this early stage of the project but several issues related to large files are present. The strategy used is to have two repositories, one for the software suite itself and one for input-files so that a user can download only the needed input files for the application he wants to run. There will only be rare cases where the whole suite will have to be downloaded, this will only be the case for the runs on the Tier-0 systems.

iRODS [44] has been evaluated but considered not suitable for WP7 basic needs (mainly upload/download of files) since it adds a lot of overhead.

For the benchmarks JuBE [45] is used to produce result tables which are manually included in the PRACE WIKI. A database backend for JuBE is under development.

6 Conclusions and future work

The work WP6 has done so far, has laid the grounds for a seamless pan-European infrastructure of Tier-0 systems with a common set of services that allow to provide a single interaction layer for users and a coordinated operational management. The infrastructure is also capable to be extended to Tier-1 systems in the near future.

Compute services are provided by both local and global mechanisms, which are a direct interaction with documented batch scheduler systems at system layer and by relying on job management functionalities of UNICORE 6 at an infrastructure layer. Data management, which is a key issue to spread more and more the use of the PRACE-RI to different scientific communities, relies on GridFTP, a standard *de-facto* on this field, and on user clients for an easy interaction. Future solutions are under evaluation on this area.

Access to resources relies on the use of X.509-based certificates for allowing both single sign-on features and security of the infrastructure.

A monitoring infrastructure has been defined and sample reports are going to be produced, together with a role-based access model. Also Network Services have as mission to provide useful information to user for tuning the communication pattern of their applications running on Tier-0 systems.

A single accounting system has been created managed and used by different entities (PRACE AISBL, PRACE Project Staff, Tier-0 Centres, Project Principal Investigators, End Users).

The defined model for provisioning of user support, and its technical implementation, reflects this view allowing users to have a single point of contact for having support even if the underlying infrastructure is a collection of different systems and institutions.

While each piece of software has been evaluated, further evaluation of the whole infrastructure is needed, in order to assess how the whole service set will perform for defined use cases. Service certification action [4.4] would facilitate the achievement of this goal as it will guarantee that all services, wherever they run, comply with user demands. WP6 will start the preparation for this evaluation; services will be tried out from a user perspective and their implementations enhanced in order to gradually satisfy technical requirements. Within this context, a refinement of the proposed solutions is expected, based on actual usage and feedback from the operations teams at Tier-0 sites.

New technologies will be evaluated to keep PRACE current with the rapid change of resource utilization models. Collaborations with other projects and initiatives, such as EMI, IGE, MAPPER, are also of vital importance to encounter user’s demands. The realization of an integrated environment requires the adoption of appropriate technological solutions to simplify the interoperability among its different components. Exchanges of experiences and solutions among e-Infrastructure projects can largely speed up the achievement of this objective. Increasing collaborations will enable a thoughtful growth of infrastructure layer enriching the support offered to our users. Neglecting this aspect, would set us back in the evolution process.

Appendix A

The aim of the following tables is to present for each core service, the list of non-functional requirement that have been identified so far. For more details on this activity please refer to Par. 4.3.2.

Service Name	Unified access to HPC infrastructure (Ref.: Services Catalogue)
Technology	GSI-SSH
Functional requirements	
• Provide users access to PRACE computational resources	
Non-functional requirements	
• High availability/reliability: the service must support high availability configuration to reduce service downtimes. • Open source and based on widely adopted standards to facilitate the interoperability among different research infrastructures (i.e. EGI, TeraGrid, etc.). • Authentication protocol should be based on Open Standards (i.e. PKI). • Authentication should be two-way: Client-to-Service, Service-to-Client. • Well-supported by developers or company. • Logging: date, time, source IP, username for any tentative access. • No credentials passing in clear. • Credentials not stored at all. • Opportunity to configure with credentials expiration, history, and intruder lockout	
Evaluation	
See section: 0	

Table 11: Unified access to HPC infrastructure.

Service Name	Data transfer, storage and sharing (Ref.: Services Catalogue)
Technology	GridFTP
Functional requirements	
• Provide users access to and from data workspace	
Non-functional requirements	
• High availability/reliability: the service must support high availability configuration to reduce service downtimes. • High efficiency in resource utilization and performance while transfers of data. • Open source and based on widely adopted standards to facilitate the interoperability among different research infrastructures.	
Evaluation	
See section 4.6	

Table 12: Data transfer, storage and sharing.

Service Name	Authentication (Ref.: Services Catalogue)
Technology	X.509
Functional requirements	
• Provide users authentication enforcement.	

Non-functional requirements
• High availability/reliability: the service must support high availability configuration to reduce service downtimes. • Easy to process at the user side – eg. Single-sign-on and without additional installations required at the user side. • Ensuring the credentials' owner, that the system is reliable and security tokens will stay secure all the time.
Evaluation
See the section 0

Table 13: Authentication.

Service Name	Authorization
Technology	LDAP
Functional requirements	
• Provide users access to PRACE resources.	
Non-functional requirements	
• High availability/reliability: the service must support high availability configuration to reduce service downtimes. • Ease to manage the resource access policies for the user/ group/ project by the operating staff.	
Evaluation	
See the section 0	

Table 14: Authorization.

Service Name	Accounting and Reporting (Ref.: Services Catalogue)
Technology	DART, Grid-Safe(not yet in production)
Functional requirements	
• Provide users access to their accounting data. • Periodic reports of system utilization from the Tier-0 sites for use by PRACE widely	
Non-functional requirements	
• Intuitive and easy of usage interfaces for users to access accounting records. • Secure, uniform and trustworthy reporting system. • Automatic notification of the events: approaching to the end of the granted time, etc.	
Evaluation	
See section: 0	

Table 15: Accounting and Reporting.

Service Name	Resources Monitoring (Ref.: Services Catalogue)
Technology	Inca
Functional requirements	
• Watches and analyzes essential PRACE parameters to keep track of the situation of the distributed RI, including: system uptime/downtime and general usage levels, network connections, incidents, software and service availability, …	
Non-functional requirements	
• High availability/reliability: the service must support high availability configuration	

to reduce service downtimes. • Easy to achieve the information on the current available subsystems: it should answer the user, whether his working environment is ready to use just after the logging into the system. • Maintain historical information on service availability. • Low impact on infrastructure performance.
Evaluation
See the section 4.10

Table 16: Resources Monitoring.

Service Name	Network Monitoring (Ref.: Services Catalogue)
Technology	perfSONAR, Iperf, bwctl
Functional requirements	
• Provide the data on the state of the links and its parameters	
Non-functional requirements	
• High availability/reliability: the service must support high availability configuration to reduce service downtimes. • Fast (on-line) network problem discovery and notification. • Easy to get the processed output from the monitoring.	
Evaluation	
The implementation was ready and/or is on the way now with four-folded approach: • Iperf tests between HPC systems. Special software (selectively PERL, C, or Python) on HPC system for managing iperf and ping tests between all partners (crontab-based, done by local PRACE partners). It has been in production in DEISA already. • Iperf tests between PerfSONAR monitoring systems PerfSONAR bwctl based iperf measurement (done by PRACE NOC). It has been tested in DEISA2, but not brought into production because of software bugs and pure performance. This will be evaluated in PRACE once more with new software. Hopefully available at the end of September 2011. We need monitoring PCs at every site (available already at many partner sites). Some components of PerfSONAR software, mainly bwctl, and iperf. A user interface to lookup monitoring data. Having such a user interface would allow to include the measured data into Inca. It is not yet clear if "users of the HPC systems" should get access to this data or only administrative PRACE staff and PRACE procedures. • Monitoring of the PRACE backbone SNMP-based, PerfSONAR E2E-monitoring, CISCOworks (done by PRACE NOC). Already implemented and accessible by PRACE NOC only. • Site-specific monitoring of the PRACE local subnets. Site dependent monitoring software (done by local PRACE partner). Already implemented and accessible by local PRACE site network staff only.	

Table 17: Network Monitoring.

Service Name	Software Management and Common Production Environment (Ref.: Services Catalogue)
Technology	‘Module’ tool

Functional requirements
• Provides software, tools, libraries, compilers, and uniform mechanisms for software and environment configuration hiding the environment complexity.
Non-functional requirements
• Easy of usage, eg. Switching compiler environment, managing environment for several dependant applications at the same time without user assistance. • Well supported be developers. • Highly adaptable to be used on as many as possible computing platforms.
Evaluation
See the section 4.9

Table 18: Software Management and Common Production Environment.

Service Name	Job processing (Compute Services)
Technology	UNICORE, local batch scheduler systems
Functional requirements	
• Provide users a fair share of the computational resources.	
Non-functional requirements	
• High availability/reliability: the service must support high availability configuration to reduce service downtimes. • Low waiting time in queues • The cause, why the job is pending at the time. • The approximated time of the placement and starting of the job. • MPI (or any other parallel/distributed) environment set automatically. • Descriptive exit codes of the queuing system. • Automatic resubmission of the job after machine crash-recovery and fair accounting of the previously exited job. • The possibility to check on-line, whether the job is running using number of cores and their utilization as much as requested; the equivalent of 'top' command on SMP machine. • E-mail notification on the specific event.	
Evaluation	
See the section 4.7	

Table 19: Job processing.