



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

D6.3

Second Annual Report on the Technical Operation and Evolution

Version: 1.0
Author(s): Gabriele Carteni (BSC), Giuseppe Fiameni (CINECA), Xavier Delaruelle (CEA), Jules Wolfrat (SARA), Axel Berg (SARA)
Date: 25.06.2012

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: D6.3	
	Deliverable Nature: DOC_TYPE: Report	
	Deliverable Level: PU	Contractual Date of Delivery: 30 / June / 2012
		Actual Date of Delivery: 30 / June / 2012
EC Project Officer: Thomas Reibe		

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

Document Control Sheet

Document	Title: Second Annual Report on the Technical Operation and Evolution	
	ID: D6.3	
	Version: 1.0	Status: Final
	Available at: http://www.prace-project.eu	
	Software Tool: Microsoft Word 2010	
	File(s): D6.3.docx	
Authorship	Written by:	Xavier Delaruelle (CEA), Gabriele Carteni (BSC), Giuseppe Fiameni (CINECA), Jules Wolfrat (SARA), Axel Berg (SARA)
	Contributors:	Thomas Boenisch (HLRS), Stephanie Meier (FZJ), Ralph Niederberger (FZJ), Michael Rambadt (FZJ), Jutta Docter (FZJ), Marcello Morgotti (CINECA), Mirosław Kupczyk (PSNC), Liz Sim (EPCC), Ilya Saverchenko (LRZ), Jarno Laitinen (CSC), Bartosz Kryza (PSNC), Giacomo Mariani (CINECA)
	Reviewed by:	Thomas Eickermann (FZJ), Nuzhet Dalfes (UYBHM)
	Approved by:	MB/Technical Board

Document Status Sheet

Version	Date	Status	Comments
0.98	10/06/20 ¹²	Draft for internal review	
1.0	25/06/2012	Final	Outline

Document Keywords

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

Disclaimer

This deliverable has been prepared by the responsible Work Package of the Project in accordance with the Consortium Agreement and the Grant Agreement n° RI-261557 . It solely reflects the opinion of the parties to such agreements on a collective basis in the context of the Project and to the extent foreseen in such agreements. Please note that even though all participants to the Project are members of PRACE AISBL, this deliverable has not been approved by the Council of PRACE AISBL and therefore does not emanate from it nor should it be considered to reflect PRACE AISBL's individual opinion.

Copyright notices

© 2012 PRACE Consortium Partners. All rights reserved. This document is a project document of the PRACE project. All contents are reserved by default and may not be disclosed to third parties without the written consent of the PRACE partners, except as mandated by the European Commission contract RI-261557 for reviewing and dissemination purposes.

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

Table of Contents

Document Keywords	ii
Table of Contents	iii
List of Figures	v
List of Tables.....	vi
References and Applicable Documents	vi
List of Acronyms and Abbreviations	vii
1 Introduction	2
2 PRACE sustainable services	3
2.1 Introduction	3
2.2 PRACE Service Catalogue	5
2.3 PRACE Operational Key Performance Indicators	6
Service availability	7
2.4 PRACE Security Forum	7
2.4.1 <i>Security Policies and Procedures</i>	7
2.4.2 <i>Risk reviews</i>	8
2.4.3 <i>Operational security</i>	8
2.5 Collaboration with other e-infrastructures.....	9
2.5.1 <i>MAPPER</i>	9
2.5.2 <i>EMI</i>	11
2.5.3 <i>IGE</i>	12
2.5.4 <i>EGI</i>	13
3 Status and planning of Tier-0 services.....	14
3.1 Technical overview of current Tier-0 production systems	14
3.1.1 <i>JUGENE – GCS@FZJ</i>	14
3.1.2 <i>CURIE – GENCI@CEA</i>	14
3.1.3 <i>HERMIT – GCS@HLRS</i>	15
3.1.4 <i>SuperMUC – GCS@LRZ</i>	17
3.2 Planning of new Tier-0 systems.....	20
3.2.1 <i>FERMI – CINECA</i>	20
3.2.2 <i>MareNostrum – BSC</i>	22
4 Selection and deployment of common services	22
4.1 Network services.....	22
4.2 Data services	22
4.2.1 <i>Status of Deployment</i>	22
4.2.2 <i>Documentation</i>	23
4.2.3 <i>Secure setup</i>	23
4.2.4 <i>Advanced User Tools</i>	24
4.2.5 <i>Monitoring</i>	24
4.3 Compute services.....	24
4.3.1 <i>Local Batch Systems</i>	24
4.3.2 <i>UNICORE</i>	25
4.4 AAA services.....	27
4.4.1 <i>Public Key Infrastructure - PKI</i>	28

4.4.2	<i>User Administration</i>	28
4.4.3	<i>Interactive access</i>	29
4.4.4	<i>Accounting services</i>	29
4.5	User services	30
4.5.1	<i>PRACE Common Production Environment</i>	30
4.5.2	<i>User Documentation</i>	31
4.5.3	<i>The PRACE Trouble Ticket System</i>	31
4.6	Monitoring services	32
5	Identification, selection and evaluation of new technologies	35
5.1	Requirement analysis	35
5.2	Service certification	36
5.3	PRACE Information System	41
5.4	Service areas	44
5.4.1	<i>Network services</i>	44
5.4.2	<i>Data services</i>	44
5.4.3	<i>Compute services</i>	49
5.4.4	<i>AAA services</i>	54
5.4.5	<i>User services</i>	56
5.4.6	<i>Monitoring services</i>	57
5.5	Summary of relevant achievements	58
6	Conclusions	59
7	Appendix A: PRACE Service Catalogue	61
	Core services	62
	Additional services	62
	Optional services	62
	Uniform access to HPC	63
	PRACE internal interactive command-line access to HPC	63
	PRACE external (user) interactive command-line access to HPC	64
	Project submission	64
	Data transfer, storage and sharing	65
	HPC Training	65
	Documentation and Knowledge Base	65
	Data Visualization	66
	Authentication	66
	Authorization	67
	Accounting	67
	Information Management	68
	Network Management	68
	Monitoring	69
	Reporting	69
	Software Management and Common Production Environment	70
	First Level User Support	70
	Advanced User Support	70
8	Appendix B: PRACE Operational Key Performance Indicators	74

Service availability	75
Service reliability	76
Number of service interruptions.....	76
Duration of service interruptions.....	77
Availability monitoring	78
Number of major security incidents	78
Number of major changes	79
Number of emergency changes.....	80
Percentage of failed release component acceptance tests.....	80
Percentage of failed service validation tests	81
Number of incidents	81
Average initial response time.....	82
Incident resolution time	82
Resolution within SLA.....	83
Number of service reviews.....	83
9 Appendix C: Sample quality checklist for Uniform Access to HPC service	85

List of Figures

Figure 1: PRACE Service provision scheme and contracts to its users.....	3
Figure 2: process towards Quality of Service in PRACE	4
Figure 3: MAPPER overall architecture.....	10
Figure 4: JUGENE	14
Figure 5: CURIE	15
Figure 6: Hermit.....	15
Figure 7: Conceptual Architecture of the Hermit system.....	16
Figure 8: Rendering of the SuperMUC installation.....	17
Figure 9: The structure of the SuperMUC system	18
Figure 10: FERMI network and storage schema	21
Figure 11: Principle Setup of the Tier-0 GridFTP at HLRS	23
Figure 12: Monitoring of GridFTP status.....	24
Figure 13: UNICORE deployment design on Tier-0 systems.....	26
Figure 14: UNICORE deployment status on Tier-0 systems.....	27
Figure 15: PRACE LDAP directory tree	29
Figure 16: Accounting architecture.....	30
Figure 17: PRACE TTS Self Service Interface.....	32
Figure 18: Service Certification state diagram	37
Figure 19: Architecture for PRACE Service Certification platform.....	40
Figure 20: Google Apps Status.....	43
Figure 21: UFTP and GridFTP transfer comparison.	45
Figure 22: Schema of the gtransfer multi path functionality	46
Figure 23: High-level logical design of the software infrastructure implemented by SysFera-DS	51
Figure 24: Project statistics page provided by SysFera-Webboard.....	52
Figure 25: Accounting architecture with Grid-SAFE facility	54
Figure 26: Grid-SAFE client interface	55
Figure 27: Usage report for a PRACE project	55
Figure 28: Budget information for some projects	56
Figure 29: PRACE Service provision scheme and contracts to its users.....	61

List of Tables

Table 1: Example of a complete operational KPI description.....	7
Table 2: Basic information of Hermit at HLRS	17
Table 3: Basic information of SuperMUC at LRZ.....	19
Table 4: Basic information of FERMI at CINECA.....	22
Table 5: Deployment Status of Tier-0 sites	23
Table 6: Batch System Inventory for Tier-0 Systems	25
Table 7: UNICORE software components deployed on Tier-0	26
Table 8: Service Certification implementation work-plan	41
Table 9: List of the PRACE information providers	42
Table 10: Basic Information Map.....	43
Table 11: Comparison table for the three tools evaluated.....	47
Table 12: Comparison table among data management functions and available technologies	49
Table 13: Evaluation outcomes for ProActive and SysFera software solutions	53
Table 14: Classification of PRACE Services as part of the PRACE Service Catalogue	62

References and Applicable Documents

- [1] PRACE Research Infrastructure AISBL: <http://www.prace-ri.eu>
- [2] PRACE-1IP D6.1 ‘Assessment of PRACE operational structure, procedures and policies’
- [3] PRACE-1IP D6.2 on First annual report on technical operation and evolution.
- [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] Globus GSI-OpenSSH: <http://www.globus.org/toolkit/security/gsiopenssh>
- [9] Usage Record – Format recommendation: <http://www.ogf.org/documents/GFD.98.pdf>
- [10] 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
- [11] DART: <http://www.deisa.eu/usersupport/user-documentation/deisa-accounting-report-tool>
- [12] Inca: <http://inca.sdsc.edu/drupal/>
- [13] Grid-SAFE: <http://gridsafe.forge.nesc.ac.uk/Documentation/GridSafeDocumentation/>
- [14] UNICORE: <http://www.unicore.eu/>
- [15] UNICORE software repository: <http://sourceforge.net/projects/unicore/files/>
- [16] EGI (European Grid Infrastructure): <http://www.egi.eu>
- [17] ProActive Parallel Suite: <http://proactive.inria.fr>
- [18] Globus Online: <https://www.globusonline.org/>
- [19] IGE project: <http://www.ige-project.eu/>
- [20] PRACE-1IP deliverable D7.4.3 Tier-0 Applications and Systems Usage
- [21] ActiveEon, <http://www.activeeon.com>
- [22] SysFera-DS, <http://www.sysfera.com>
- [23] DIET, Distributed Interactive Engineering Toolbox, <http://graal.ens-lyon.fr/DIET>
- [24] Decryphon Grid, <http://www.decryphon.fr/>
- [25] PRACE Helpdesk, <http://tts.prace-ri.eu>
- [26] EGI GGUS Helpdesk, <http://helpdesk.egi.eu/>

- [27] XSEDE, <https://www.xsede.org/>
- [28] MAPPER, <http://www.mapper-project.eu>
- [29] PRACE Tier-0 infrastructure utilization scenario
- [30] BSC MareNostrum, Usage Status, <http://www.bsc.es/marenostrum-support-services>
- [31] EPCC HECToR Usage Status, <http://www.hector.ac.uk/service/status/>
- [32] FZJ Maintenance Messages, http://www2.fz-juelich.de/jsc/CompServ/services/high_msg.html
- [33] BSC MareNostrum, Available Software, <http://www.bsc.es/marenostrum-support-services/available-software>
- [34] HECToR Software versions, <http://www.hector.ac.uk/service/software/>
- [35] INCA web pages, <http://inca.prace-ri.eu>
- [36] PRACE Network Information Pages, <http://net.prace-ri.eu>
- [37] PRACE Data Management White Paper, <https://bscw.zam.kfa-juelich.de/bscw/bscw.cgi/d730117/White%20Paper%20on%20Data%20Management.pdf>
- [38] The Grid-SAFE project, <http://www.epcc.ed.ac.uk/projects/grid-safe>
- [39] XSEDE User Portal, Resource Monitor, <https://portal.xsede.org/resource-monitor>
- [40] NERSC (National Energy Research Scientific Computing Center), Resource Status, <http://www.nersc.gov/users/live-status/>
- [41] Reference DECI user feedback collected during DEISA or the respective reports/deliverables
- [42] Globus Toolkit, <http://www.globus.org/toolkit/>
- [43] Gtransfer, <https://github.com/fr4nk5ch31n3r/gtransfer>
- [44] MAPPER project deliverables, <http://www.mapper-project.eu/web/guest/documents>

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 (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)
CSIRT	Computer Security Incident Response Team
DART	Distributed Accounting Report Tool
DCI	Distributed Computing Infrastructure
DECI	Distributed Extreme Computing Initiative
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 ($= 2^{30} \sim 10^9$) 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 ($= 2^{20} \sim 10^6$) 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
RRD	Round-robin database
SAML	Security Assertion Markup Language
SCI	Security for Collaborating Infrastructures
SNIC	Swedish National Infrastructure for Computing (Sweden)
SPG	Security Policy Group
SRM	Storage Resource Management

D6.3 Second Annual Report on the Technical Operation and Evolution

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

The objective of this report is to present the work that has been done in year 2 of the PRACE-1IP project work package 6 on the Operations of the PRACE distributed infrastructure.

To support a good and complete overview, description and classification of PRACE Operational services, we have developed a PRACE Service Catalogue last year. In year 2 Tier-1 services have been added to this Service Catalogue to complete the picture of PRACE service provision. Further, the PRACE Service Catalogue has been fine-grained by naming and classifying for every service also the actual product that is used to provide the service. Every product has been classified as well. In this manner, a complete overview of all PRACE services, products and their classification has been made.

In the process towards PRACE Quality of Service and quality control we have started defining PRACE Operational Key Performance Indicators. A first set of KPIs have been developed and proposed by a working group and these are currently discussed among the operational partners. In parallel, the measurement of the KPIs is being implemented.

Based on the procedures for incident and change management that have been setup in year 1, we have operated a complete set of PRACE common services in year 2. 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).

A number of Tier-0 systems have been installed and integrated in PRACE Operations in year two:

- an upgraded system of CURIE from GENCI@CEA;
- the new system HERMIT from GCS@HLRS;
- and the first nodes of SuperMUC from GCS@LRZ

Preparations have been made for the integration of the FERMI system at CINECA.

Collaborations on the operational level have been continued with EGI, EMI, IGE and MAPPER.

On the technical evolution of the common services of the infrastructure, we have consolidated and extended existing services evaluating new technologies on the basis of requirements from users that have been collected. Moreover, we have improved the level of offered services through the definition of a service certification process. We have further extended the accounting systems to provide more information about allocated resource budgets, and we have laid a foundation for the development of a data management strategy.

1 Introduction

The presentation and delivery of the PRACE services to its users as a single coordinated distributed research infrastructure allows users to use the PRACE infrastructure [1] as seamlessly as possible. To establish and assure this, coordinated actions are required on many different levels, from peer review and training activities to service deployment around the actual use of the infrastructure. Work package six (WP6) deals with the coordination of the technical operations and the technical evolution of the distributed PRACE infrastructure.

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, making the PRACE infrastructure more than the sum of its individual centres, systems and services. To achieve this goal, the operational work has been divided and organised along three main tasks:

- (T6.1) Establishment of an organizational structure coordinating the technical operations: work on organizational structure, service catalogue, operational procedures (incident & change management), model for user support, key performance indicators
- (T6.2) Provision, operation and integration of comprehensive common services at the Tier-0/1 level: work on running the operational coordination and deployment of common services (compute, data, network, user, AAA, monitoring, generic)
- (T6.3) Technical evolution of the distributed Research Infrastructure: work on requirements analysis (users, technical), technology watch and assessment, selection and testing of services, service certification

This report describes the work that has been done in WP6 in the second and final project year of PRACE-IIP, on the technical operation and evolution of the infrastructure. The results of year 1 of WP6 have been described in two deliverables, D6.1 on Assessment of PRACE operational structure, procedures and policies [2], and D6.2 on First annual report on technical operation and evolution [3].

The structure of this report reflects the organization of the work package and the outcomes produced by its (sub)tasks:

- Chapter 2 is on PRACE sustainable services and provides the roadmap and results with respect to PRACE sustainable quality of service, including a description of the work on the PRACE Service Catalogue, PRACE Operational Key Performance Indicators, and collaboration with other e-infrastructures. These activities have been performed within task T6.1.
- Chapter 3 summarizes the status and planning of the actual Tier-0 services, and gives a technical overview of the Tier-0 production systems and provides a planning of the deployment of new Tier-0 systems.
- Chapter 4 summaries all activities carried out within Task 6.2 during the second year on the deployment of PRACE common services. All information is organized into sub-sections, each one reflecting the work done within each service category.
- Chapter 5 is dedicated to Task 6.3 and reports on assessment of new technologies following the same structure as Chapter 4.
- Chapter 6 reports some final conclusions.

2 PRACE sustainable services

2.1 Introduction

The PRACE distributed research infrastructure is operated and presented to the users as a single research infrastructure, allowing the users to use PRACE as seamlessly as possible. This is done by Tier-0 hosting partners working closely together and synchronising service provision and service deployment as much as possible. PRACE common services are deployed that provide a service layer that integrates the various hosting partner Tier-0 services, and makes the PRACE infrastructure much more than just a collection of individual Tier-0 hosting partners and Tier-0 services.

The PRACE distributed research infrastructure is well on its path to provide a complete set of sustainable services to its users. Service provision to users is currently mainly done by the Tier-0 hosting partners, governed by the PRACE AISBL statutes and the Agreement for the Initial Period. Relations between Tier-0 sites and their users are typically managed through specific User Agreements. PRACE AISBL gives advice to the hosting sites on the allocation of compute resources based on the pan-European PRACE Peer Review. For the execution of the peer review and other services such as the PRACE website, the PRACE also uses services provided by third parties. Other important services such as user support and operation of the distributed infrastructure are provided by the PRACE-IIP project.

Tier-1 partners provide access to users, governed by the DECI commitments, currently within the Implementation Phase projects.

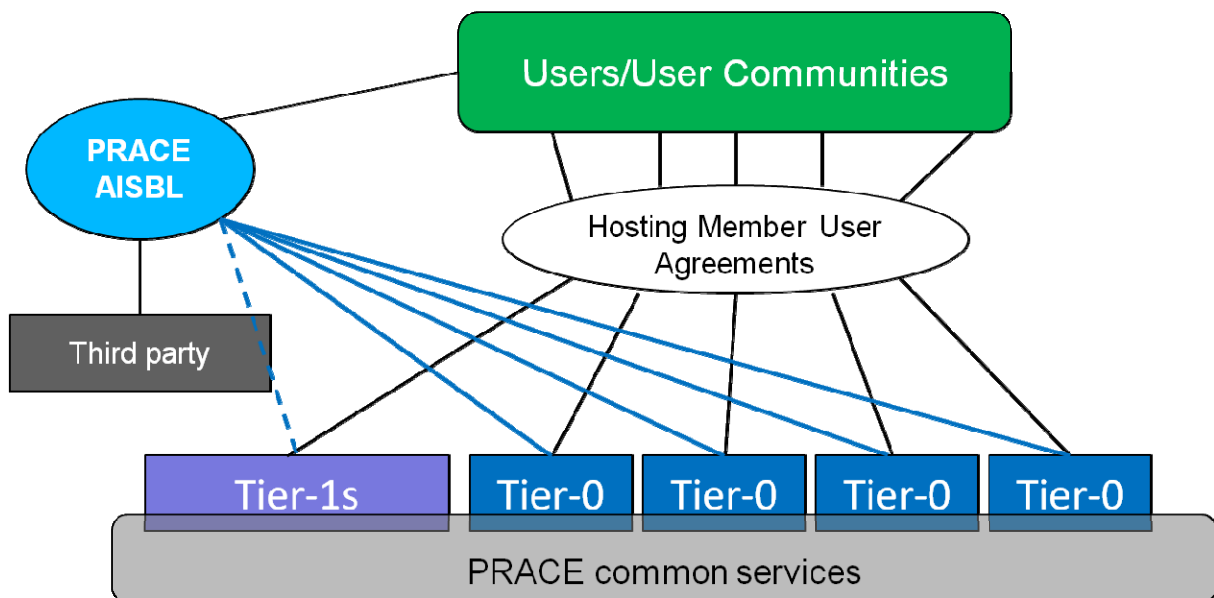


Figure 1: PRACE Service provision scheme and contracts to its users

In the process towards the provision of sustainable and reliable PRACE common services of defined and professional quality, we have made at the start of the project a clear roadmap with distinct steps to achieve quality of service on the long term. This roadmap is illustrated by Figure 2 and contains the following steps:

1. **the definition and agreement of the set of PRACE common services:** in year 1 we have created a first version of the **PRACE Service Catalogue**, in this reporting year 2 we have refined the PRACE Service Catalogue and added also Tier-1 services. The PRACE Service Catalogue describes the PRACE common services, as well as their service classes (core, additional, optional);
2. **the definition and implementation of the PRACE Operational Structure** through the PRACE Operational Coordination Team, at the start of year 1 in a matrix organisation with site representatives and service category leaders;
3. the definition and **implementation of a model for user support:** in year 1 we have setup a central helpdesk that is locally managed;
4. the definition, agreement and **implementation of operational procedures and policies** for the service delivery: in year 1 we have described and implemented common procedures for incident and change management;
5. the definition of a **service certification** process to verify, ensure, control and improve the quality of services to be deployed newly; in year 1 and 2 we have defined a complete process for service certification;
6. the definition of a starting set of **operational Key Performance Indicators (KPIs)**: in year 2 we have proposed a set of operational KPIs that are currently being implemented;
7. the measurement of KPIs followed by the **definition of service levels** for each of the services: this activity is to be taken up by the Operations work package of the PRACE 2IP and 3IP project.

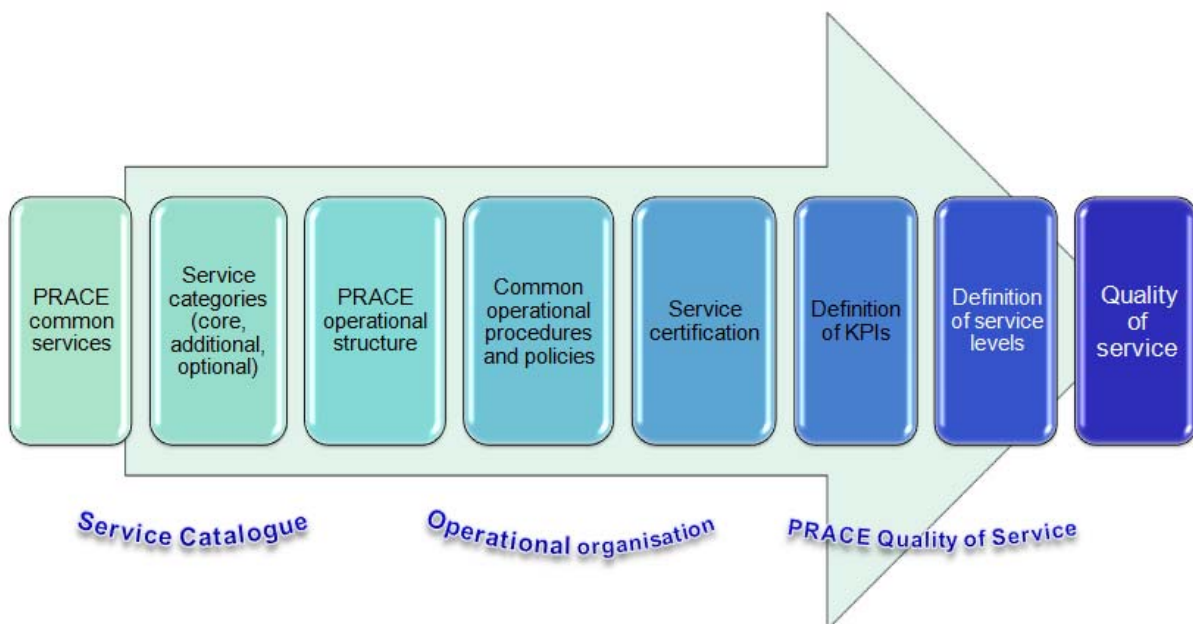


Figure 2: process towards Quality of Service in PRACE

All these steps are a prerequisite for the implementation of a sustainable set of PRACE common services with quality assurance and quality control (see also 2.3 on PRACE Operational Key Performance Indicators).

During the course of this project we have made significant progress on the implementation towards Quality of Service. In year 2 the work was mainly dedicated to the update and fine-graining of the PRACE Service catalogue (see paragraph 2.2) and to the definition and implementation of a first draft set of Operational Key Performance Indicators (see paragraph 2.3).

2.2 PRACE Service Catalogue

To support a good and complete overview of all PRACE Operational Services, we have developed the PRACE Service Catalogue, which lists and describes the complete set of operational services that the PRACE AISBL is providing, from the point of view of PRACE as a service provider.

The purpose of the PRACE Service Catalogue is:

- To describe all PRACE operational services
- To define PRACE service categories, and classify all PRACE services accordingly

In this way it describes the full PRACE service portfolio from hosting partners, other partners, the project and the PRACE AISBL.

An important aspect of the PRACE Service Catalogue is the classification of services. We have defined three service classes: Core services, Additional services and Optional services. The details of this classification can be found in the current version of the PRACE Service Catalogue in Appendix A.

Every PRACE service will be classified according to this classification. It should be noted that the service classes define the availability of the services at the hosting sites, and are not related to service levels.

The PRACE Service Catalogue is regularly updated to document the actual status of all services and will be maintained as a living document, where all changes in services and their provision will be indicated. Status of services can change when new services are deployed, when levels of services are changed, when new service providers (i.e. new hosting partners) are integrated or when new software products are released. The document will at all times reflect the current situation of PRACE services, so that it can be used as the main reference document for service provision within PRACE.

The starting point for the list of services that is listed in the PRACE Service Catalogue has been established already in the PRACE-PP in WP4.

In year 1 of the PRACE-1IP project we have started to develop the PRACE Service Catalogue for PRACE Tier-0 services, as described in deliverable D6.1. In year 2 Tier-1 services have been added to this Service Catalogue to complete the picture of PRACE service provision. Further, the PRACE Service Catalogue has been fine-grained by naming and classifying for every service also the actual product that is used to provide the service. Every product has been classified as well. In this manner, a complete overview of all PRACE services, products and their classification has been made.

The first version of the PRACE Service Catalogue has been sent for feedback to the PRACE Board of Directors in September 2011. The current version of the PRACE Service Catalogue has been presented to the PRACE Board of Directors for feedback and approval by the end of March 2012. Awaiting feedback and approval by the PRACE BoD the PRACE Management Board has decided that the project should meanwhile act as if the PRACE Service Catalogue was approved. After that, the PRACE Service Catalogue will be submitted to the User Forum for comments.

The complete current version of the PRACE Service Catalogue (v1.6a) can be found in Appendix A.

2.3 PRACE Operational Key Performance Indicators

Quality assurance and quality control are important whenever services are delivered. In the PRACE situation the service delivery is complex as it is delivered as a 'single' PRACE service to the users, but actual service delivery is a combination of services provided by many hosting partners and other partners. Quality assurance is the systematic monitoring and evaluation of services to maximize the probability that service levels are being attained by the service delivery process.

In the process towards quality of service as described in paragraph 2.1 we have drafted a first and limited set of PRACE Operational Key Performance Indicators (KPIs).

The objective of the operational KPIs is to provide insight in how well we are doing on PRACE Operations, based on facts (measurable), and based on expectations (what is the level of service that we consider satisfactory). In our approach we took a stepwise approach:

- Define a limited number of Operational Key Performance Indicators (KPIs):
 - o Measurable quantitative values
 - o Based on ITIL categories
 - o Give information on the quality of the service & service provision
 - o Periodically monitored and registered
- Implement measurement/monitoring of the KPIs
- Measure for a certain period of time to determine what is a realistic regular_level of the KPI
- Define the Service Level → Measure against this level = Quality Control

We have drafted 14 Operational KPIs (based on ITIL Categories):

- (1) Service Availability
- (2) Service Reliability
- (3) Number of Service Interruptions
- (4) Duration of Service Interruptions
- (5) Availability Monitoring
- (6) Number of Major Security Incidents
- (7) Number of Major Changes
- (8) Number of Emergency Changes
- (9) Percentage of Failed Services Validation Tests
- (10) Number of Incidents
- (11) Average Initial Response Time
- (12) Incident Resolution Time
- (13) Resolution within SLA
- (14) Number of Service Reviews

For each of these KPIs, we have defined:

- Description
- Calculation
- Inputs
- Outputs
- Time-interval for measurement
- Tools for measuring the KPI
- ITIL Category for reference
- Implementation plan

An example of such KPI description is given in table 1 below for the Service availability KPI.

Service availability	
Description:	Availability of services
Calculation:	$((A-B) / A) * 100$
Inputs:	Committed hours of availability (A) Outage hours excluding scheduled maintenance (B)
Outputs:	Availability (%)
Time-interval:	Bi-weekly (during every PRACE Operations meeting)
Threshold:	
Tools:	Inca
ITIL Category:	Service Design – Availability Management
Implementation plan:	<p>Inca provides all data necessary for computing this KPI. All test results for a specific service over a given period of time have to be extracted from Inca. Based on the extracted data the total number of tests and the number of failed tests has to be computed. These two numbers should be used in the formula above to compute service availability.</p> <p>The necessary data can be extracted and processed using an SQL query and presented in the PRACE information portal.</p>

Table 1: Example of a complete operational KPI description

In Appendix B descriptions of all 14 KPIs can be found.

The status of this work is that a first set of KPIs have been developed and proposed by a working group and these are currently discussed among the operational partners. In parallel, the measurement of the KPIs is being implemented according to the proposed implementation plan. We expect agreement on a first set of operational KPIs among the operational partners and their implementation in the second half of 2012.

2.4 PRACE Security Forum

The establishment of the PRACE Security Forum was accepted at the end of the PRACE-PP project. The implementation started in the summer of 2010 with the acceptance by the PRACE-TB of a document which describes the objectives, the tasks and the organisation of this body.

The Security Forum has three main tasks:

- Defining security related Policy and Procedures;
- The Risk Review of new services or the service upgrades;
- The management of operational security.

2.4.1 Security Policies and Procedures

The main activity this year has been the discussion of security policies in the SCI (Security for Collaborating Infrastructures) working group, a collaboration of large infrastructures, currently including EGI, OSG, PRACE, WLCG, and XSEDE. SCI is developing a framework

to enable interoperation of collaborating Grids with the aim of managing cross-Grid operational security risks and to build trust and develop policy standards for collaboration especially in cases where we cannot just share identical security policy documents. PRACE security forum members participated in several video conferences and one face to face meeting to discuss a document which describes the requirements for collaborating infrastructures for 1) Operational Security; 2) Establishing Trust between Collaborating Infrastructures; 3) Participant Responsibilities; 4) Legal Issues; 5) Data Protection.

The document is in a final state and can be used internally to update or to further implement policies and procedures. A common framework will enable an easy sharing of resources from different infrastructures by user communities or projects.

Several security forum members also participate in the Security Policy Group (SPG) of EGI as non voting members. This enables the exchange of information on policy and procedures. In the past this resulted in a shared AUP for users.

PRACE is a Relying Party member of EUGridPMA [5]. One EUGridPMA meeting (Karlsruhe, May 2012) has been attended this project year by the PRACE representative. The involvement is important for monitoring the accreditation process of new CAs and the auditing of already accredited CAs. Feedback on problems and requirements is also given.

2.4.2 *Risk reviews*

The security forum must make a risk assessment of any new service or the update of an existing service if in case of the latter there are changes in the security set-up.

This period much effort has been put in the risk review of Globus Online [18]. Globus Online is evaluated as a new data management service by T6.3, but because of the high security impact it was important to have a risk review before finalizing the evaluation. Several video conferences were held to assess the risks and as a result a report was published with a negative advice because one of the risks was considered too high. The report was also communicated to the Globus Online team, using the IGE project [19] as the contact organization. As a result the Globus Online team proposed several improvements which for the security forum were reason to change their advice from negative to positive, provided that the proposed changes are implemented. These changes were further discussed in a face to face meeting of several PRACE members with a representative of Globus Online in Munich, March 2012. This resulted in a final implementation plan of the changes by the Globus Online team. As a result of this the security forum advised T6.3 to finalize the evaluation of Globus Online. The results of the risk assessment were also exchanged with the XSEDE project, which also sees Globus Online as an interesting service for their community. Results of their assessment also will be made available to PRACE.

If the Globus Online service will be added as a PRACE recommended service a contract between PRACE and Globus Online is needed to document the requirements that PRACE has with respect to security and service levels. This is needed because Globus Online is not only delivering software but also a service which is hosted in the cloud. To prepare for such a contract it is proposed to start with a MoU in the initial phase.

The experience of the risk assessment of the Globus Online service is used to publish a first draft of the risk review procedure.

2.4.3 *Operational security*

All Tier-1 sites are added as member of the PRACE CSIRT team in the fall of 2011.

No security incidents have been reported in the PRACE infrastructure.

There is a close collaboration between the PRACE and EGI CSIRT. The PRACE CSIRT team adopted the same guidelines as EGI for the distribution of information about security incidents. Information about incidents in the EGI infrastructure is distributed to the PRACE CSIRT team. Information about incidents in the PRACE infrastructure will be forwarded to EGI too if appropriate. Members of the EGI CSIRT team are subscribed to the PRACE CSIRT distribution list.

2.5 Collaboration with other e-infrastructures

In the second year of the project there has been a significant growth of interest from other projects in collaborating with PRACE, either to integrate/access its computational resources or provide support for deployed software components, such as UNICORE. Collaborations were the result of meetings with the coordinators and representatives of interested projects during conferences or events. These activities were initiated to better support user communities, to strengthen the collaboration with external technology providers, and to address PRACE technological requirements. The establishment of cooperation with other projects was carried out in parallel with the implementation of the awareness raising and dissemination campaign. The objective of the following sections is to give an overview of the three major collaborations that were set up and/or continued during the reporting period.

2.5.1 MAPPER

The MAPPER project (Multiscale APplications on EuRopean e-infrastructures) [28] aims to deploy a computational science environment for distributed multi-scale computing, on and across European e-Infrastructures, including PRACE and EGI. The collaboration between the two projects initiated in May 2011 and was coordinated via a Task Force comprising specialists from each of the three organisations (MAPPER, PRACE, EGI-Inspire). On the PRACE side, SARA, CINECA, BSC, LRZ and EPCC were involved to provide support on the technological and application area. The primary goal of this collaboration was to demonstrate the possibility to execute MAPPER multi-scale applications across PRACE and EGI resources simultaneously and, successively, extend the test-bed towards a more sustainable and persistent solution. Two applications in the fields of “stent restenosis” and “nano material science” were selected and MAPPER middleware components, necessary for making the infrastructures interact with each other, were successfully evaluated. Although the MAPPER intention is to reuse existing services, such as UNICORE and GridFTP for PRACE, it was indispensable to evaluate and deploy new software components as new specific functionalities were needed. The diagram in Figure 3 presents the MAPPER overall architecture, including its main software components.

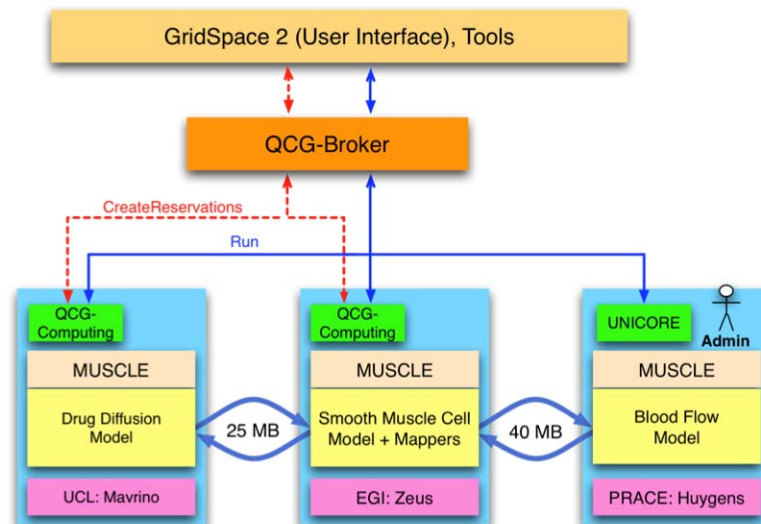


Figure 3: MAPPER overall architecture

The SARA team, who also offered their computational resources for running the demonstration, evaluated the following components and deployed them on the Dutch national supercomputer Huygens, which is part of the PRACE infrastructure for DECI. Mapper components were deployed only for testing purpose and were offered to a restricted group of users.

In order to successfully execute multi-scale applications, an advance reservation system to automatically reserve resources across sites was needed. The reason for this functionality is to optimally synchronize the runtime of the simulations with the duration of the reservations, and thereby prevent resources from idling. As most of PRACE sites, including SARA, do not support this functionality, which is also in contrast with the resource utilization policy of many HPC centres around the world, different solutions were investigated to overcome this barrier. The final solution culminated in a manual interaction of a system administrator at SARA any time an advance reservation was requested.

Results of the first prototype were positively presented at the second MAPPER official review (November 2011) where the execution of selected applications was successfully demonstrated on involved resources (UCL-Mavrino, EGI-Zeus, PRACE-Huygens).

The task force is now moving to its second phase with the intention of:

- enforcing the collaboration among involved parties, including EGI. The possibility to integrate PRACE and EGI services will be discussed;
- strengthening the collaboration between PRACE and MAPPER on the resource allocation model;
- extending the prototype including Tier-0 resources. LRZ has already manifested their interested in offering computation resources to this end;
- investigating a solution for the automatic reservation of resources.

A number of concrete requirements have been defined on a number of different topics from the MAPPER project. Actions from PRACE and EGI are currently discussed to respond to those requirements from MAPPER which are listed below:

- Resource allocation:

- *Requirements:* need for streamlined access to e-infrastructure services and resources and common mechanisms for resource allocation.

- **Resource access:**
 - *Requirements:* advance reservation and co-allocation mechanisms should be enabled on EGI and PRACE sites to support novel use-cases (see MAPPER user communities and their requirements defined in D4.1 [44]). Example: In-stent Restenosis 3D requires co-allocated resources to run at three sites concurrently (demonstrated during the first MAPPER review, Nov 2011).
- **Monitoring:**
 - *Requirements:* Statistics concerning resources availability, storage and network parameters (bandwidth and latency) should be provided to users. The middleware should offer live monitoring of simulation progress and application performance. Users should be able to define custom test probes, as what they need may not be covered by existing monitoring test suites. 3 points where monitoring info is needed: (1) before a workload is submitted for resource and service discovery, (2) during workload execution for status of the running workload and (3) after execution of the workload for analysis of running of the workload (job statistics). Need to clarify specific monitoring requirements for MAPPER jobs.
- **Accounting:**
 - *Requirements:* User should be able to monitor their allocation usage in a single federated portal. The resources usage information should be updated on daily basis. "To reduce the bureaucratic overhead of EU projects in general, and MAPPER in particular, the procedure of requesting compute time and storage must be greatly streamlined. This can be accomplished by including requests for compute time and storage space in EU project proposals." (see MAPPER deliverable D3.1, [44])
- **Security:**
 - *Requirements:* Every site should be capable of authenticating any EUgridPMA certificate (done by both PRACE and EGI). The process of acquiring X.509 credentials should be more automated and simplified.
- **User support:**
 - *Requirements:* End users should have a single point of contact for both EGI and PRACE infrastructures, as contacting each site independently is far too inconvenient. Need for a common knowledge base and set of good practices for both end users and 1-st line support. Need for help with optimization of applications for multi-cluster/site simulations (e.g. distributed multi-scale simulations)
- **Relevant MAPPER deliverables [44]**
 - D3.1 "Report on the policy framework resource providers need to adopt to support the MAPPER Project",
 - D4.1 "Review on applications, users, software and e-Infrastructures",
 - D6.3 "Support Process Definition".

2.5.2 EMI

The EMI (European Middleware Initiative) project is a close collaboration of the four major European middleware providers, ARC, dCache, gLite and UNICORE. Its aim is to deliver a consolidated set of middleware components for deployment in EGI, PRACE and other DCIs, extend the interoperability and integration between grids and other computing infrastructure. The collaboration with the EMI project initiated on September 2011 to define a common

framework of collaboration where to exchange expertise, support the evolution of UNICORE components, access to emerging technologies, enforce the sustainability of adopted technology. Both projects contribute to enable the vision of providing European scientists and international collaboration for sustainable distributed computing services to support their work. In this broad context, the specific goals of the collaborations were defined:

- to provide robust, well-designed, user-centric services to scientific user communities;
- to define a common understanding for third-level support on incidents and problems;
- to guarantee the development of specific open standards which are necessary to enable the interoperability with other e-infrastructures (e.g. the XSEDE infrastructure [27]);
- to disseminate the results of this collaboration.

A joint work-plan to implement collaboration's objectives was defined in a Memorandum of Understanding (MoU) which is currently under discussion within respective coordination bodies. The PRACE AISBL will sign the MoU if a consensus on defined objectives is reached. The MoU includes the following activities:

- **Exploitation of EMI components**
 - to permit PRACE access to EMI releases so to evaluate and test its components;
- **Evolution of UNICORE components to foster the interoperability with other e-infrastructures (e.g. XSEDE)**
 - to permit PRACE establish and maintain interoperability with XSEDE [27] through UNICORE open standards BES, JSDL, HPC-BP, SAML;
- **Evolution of the EMI Security Token Service (STS)**
 - to permit PRACE submits its requirements to support EMI in the development of the EMI Security Token Service (STS) and exploit successive results;
- **Operational Support**
 - to agree with PRACE on a support model for deployed components
- **Dissemination and Training**
 - to develop a training strategy for services which are deployed in PRACE.

Currently, the MoU is under evaluation by PRACE PMO and, after approval and signature by the PRACE AISBL, it will be implemented within the WP10 of PRACE-2IP project.

2.5.3 IGE

The Initiative for Globus in Europe (IGE) [19] is a project supporting the European computing infrastructures by providing a central point of contact in Europe for the development, customisation, provisioning, support, and maintenance of components of the Globus Toolkit [42], including GridFTP and GSI-SSH which are currently deployed in PRACE. Their support was fundamental during the security assessment of the GlobusOnline service (see paragraph 2.4) as they provided the expertise and knowledge to understand the technologies and mechanisms behind the service. Currently, the collaboration between PRACE and IGE has not been formalized yet, mainly due to lack of effort, but it will be taken forward as part of the PRACE-2IP WP10 work-plan.

2.5.4 EGI

In this second year of the project, collaboration on operations between PRACE and EGI has been intensified. The collaboration has been mainly concentrated around the MAPPER project, since MAPPER provides a number of significant use cases for collaboration and interoperation between PRACE and EGI. These use cases require within one application the use of the PRACE and the EGI infrastructure. Details on this collaboration are described in paragraph 2.5.1 on MAPPER.

During the EGI Community Forum at 26-30 March 2012 in Munich, an EGI PRACE workshop has been held, where Operational models and other details were presented and discussed.

3 Status and planning of Tier-0 services

This chapter provides a technical overview of the PRACE Tier-0 systems currently in production and available to the users. Some details are also provided on the upcoming Tier-0 systems.

3.1 Technical overview of current Tier-0 production systems

The scope of this section is to provide an update on the two Tier-0 systems already available last year and a set of technical information about the two Tier-0 systems that were integrated since then.

3.1.1 *JUGENE – GCS@FZJ*

The IBM BlueGene/P system (72 racks), named JUGENE and managed by the Jülich Supercomputing Centre, has been installed in 2009 and was first offered to the PRACE community in the summer of 2010.



Figure 4: JUGENE

The configuration of the system has been unchanged since then. Minimal required software updates and a hardware, which has proofed to run extremely stable for the last year of operation, enabled PRACE users to run jobs up to the size of the full machine.

The central fileserver JUST (Jülich Storage Server) providing filesystems via IBM's General Parallel Filesystem (GPFS) has been moved to a new server basis (running Linux) and expanded its capacities for \$WORK (4,2 PB), \$HOME (1,2 PB) and \$ARCH (600 TB) filesystems.

3.1.2 *CURIE – GENCI@CEA*

CURIE is the second PRACE Tier-0 petascale system, provisionned by GENCI and operated at TGCC (Très Grand Centre de Calcul du CEA) near Paris. CURIE has been opened to users on May 2011 and since last year two new types of hardware have been added to the supercomputer to complete the initial 90 Intel Nehalem-EX large nodes:

- 5040 Bullx B510 thin nodes with 2 Intel SandyBridge processors at 2.7 GHz (8 cores each), 64 GB of DDR3 memory (4 GB/core), SSD local disk;

- 144 Bullx B505 hybrid nodes with 2 Intel Westmere (4 cores each) 2.67 Ghz, 2 Nvidia M2090 GPUs, SSD local disk.

Along to these compute nodes addition, the interconnect network has been upgraded to a full fat tree InfiniBand QDR network. Also, the internal Lustre file system has been extended to a capacity of 5 PB and a bandwidth of 150 GB/s and computing center-wide Lustre file-system has been extended to a capacity of 8 PB and a bandwidth of 150 GB/s.



Figure 5: CURIE

3.1.3 HERMIT – GCS@HLRS

Hermit is the new Petascale System located and operated at HLRS in Stuttgart, Germany. It is the second Tier-0 system provided by the German Gauss Centre for Supercomputing for PRACE. Delivered in October 2011, Hermit has been the first Petaflop/s system world wide delivered with AMD Interlagos processors.



Figure 6: Hermit

Hermit provides a homogenous architecture over all 113,664 compute cores. All 3,552 nodes are connected to the Cray Gemini 3D-Torus network with similar network access conditions. Therefore, an application can make use of all nodes and finds similar network rates independent of the node location. Hermit is targeting for grand challenge applications which make use of the whole system or at least large partitions of it in one computational job.

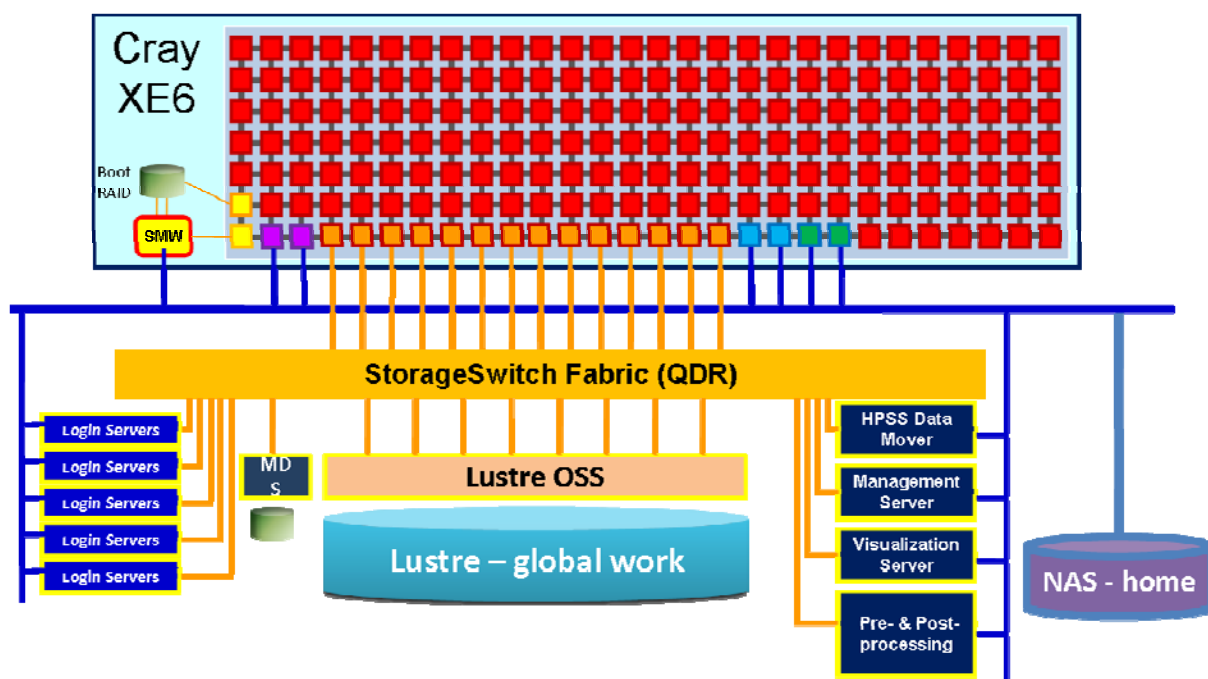


Figure 7: Conceptual Architecture of the Hermit system

Hermit provides a maximum performance of 1.045 Petaflop/s and a Linpack performance of 831.4 TeraFlop/s. The average power consumption is 1.55 MW. In addition to the hardware, Cray provides a scalable software environment with a special Cluster Compatibility Mode (CCM) which allows ISV software to run easily on the unique architecture. More technical details are available in the following table.

Machine name	Hermit1
PRACE partner	GCS@HLRS
Country	Germany
Organisation	HLRS
Location	HLRS, Stuttgart, Germany
Nature (dedicated system, access to system, hybrid)	Access to production system
Vendor/integrator	Cray
Architecture	Cray XE6
CPU (vendor/type/clock speed)	AMD / Opteron 6276 (Interlagos) / 2.3 GHz
CPU cache sizes (L1, L2, L3)	L1: 48 kB per core L2: 1 MB per core L3: 16 MB shared
Number of nodes	3552 (compute), 96 (service)
Number of cores	113664 (compute)
Number of cores per node	32 (compute)

Memory size per node	32 GB/ 64GB
Interconnect type / topology	Cray Gemini / 3D Torus
Peak performance	1.045 PFlop/s
Linpack performance (measured or expected)	831.40 TFlop/s
I/O sub system (type and size)	Home files system on a BlueArc NAS (60 TB) connected with 10 GBit Ethernet Work file system Lustre 2.7 PB connected with Infiniband; 150 GB/s bandwidth
File systems (name, type)	/zhome (HOME file system), /univ_<x> (Work file systems)
Date available for PRACE production runs	November 2011
Link to the site's system documentation	http://www.hlrs.de/systems/platforms/cray-xe6-hermit/

Table 2: Basic information of Hermit at HLRS

3.1.4 SuperMUC – GCS@LRZ

SuperMUC is the new IBM Petascale system at the Leibniz Supercomputing Centre (LRZ) in Garching (Germany). It will be available as a Tier-0 PRACE machine and is one of the top-level resources of the German Gauss Centre for Supercomputing.



Figure 8: Rendering of the SuperMUC installation.

The final configuration, with more than 150,000 cores, will deliver a peak performance of 3 Petaflops, constituting a substantial upgrade of the previous LRZ facilities.

SuperMUC consists of a Fat Node Island and 18 Thin Node Islands. The Fat Node Island is based on Intel Xeon Westmere-EX, each node being interconnected by an Infiniband QDR network. Thin Node Islands employ Intel Xeon Sandy Bridge-EP CPUs while the network technology is Infiniband FDR10. An overlay network will guarantee intra-island communications. The complete system will be fully operational by August 2012, but the Fat Node Island has already been deployed in 2011 as SuperMUC Migration System, or SuperMIG.

Users' home folders are hosted on a 1.5 PB NAS, capable of accessing data at 10 GB/s and placed on a 80 Gbits/s Infiniband trunk. GPFS was chosen for work and scratch directories, with a total space of 10 Petabytes, Input/Output performances up to 200 GB/s and served by Infiniband. Finally, 30 PB have been reserved for backups, accessible via a 10 Gigabit Ethernet connection. It will be transparent to the user, based on tapes together with a disk cache.

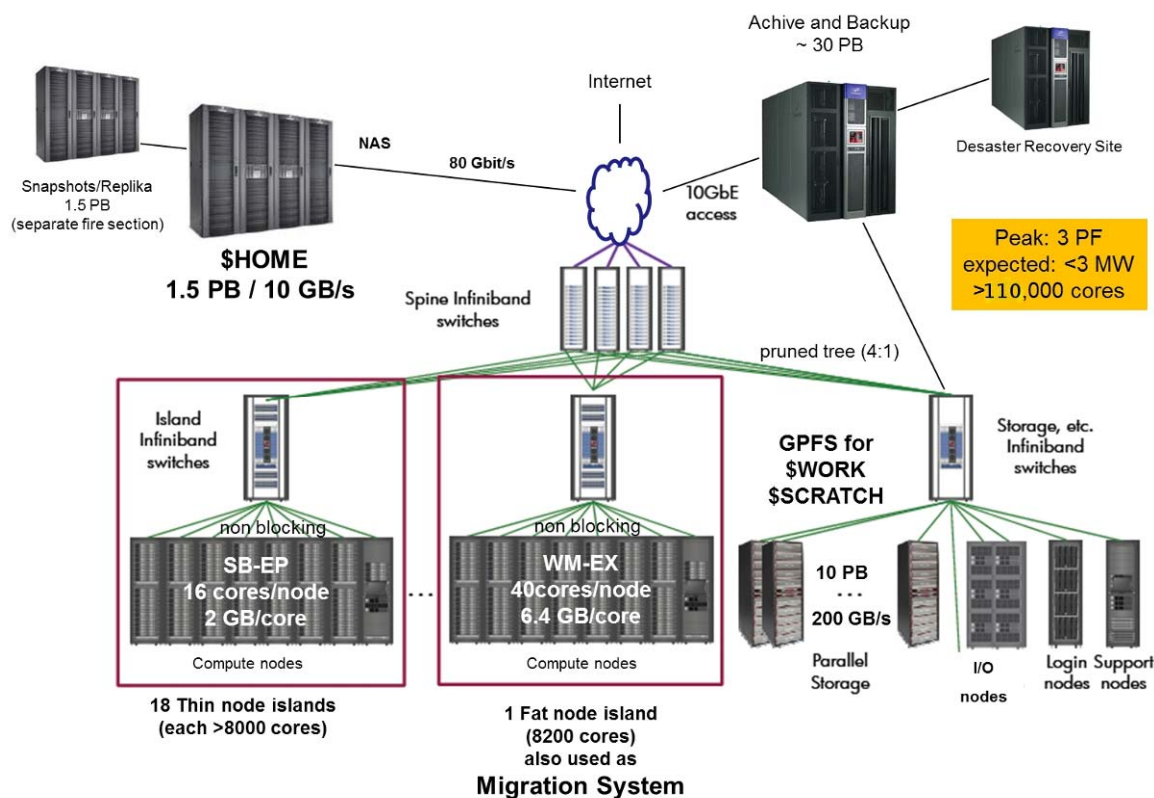


Figure 9: The structure of the SuperMUC system

SuperMUC aims also at energy efficiency, so, together with state of the art components, a new cooling system is employed. CPUs and memory are cooled by warm water: a temperature of the liquid up to 45 °C allows the system not to exceed the 80 °C limit set by the developers. About 90% of the system adopts this technology, leading to a total expected power consumption below 3 MW. Energy saving is about 40% with respect to the traditional air cooling approach. The high temperature liquid cooling system designed by IBM is known as Aquasar and SuperMUC will be one of first machines in Europe to benefit from it.

IBM is also providing the software solutions for batch job management and archiving by means of LoadLeveler and Tivoli Storage Manager, respectively.

Machine name	SuperMUC
PRACE partner	LRZ (GCS)
Country	Germany
Organisation	LRZ
Location	LRZ, Garching / Munich, Germany
Nature (dedicated system, access to system, hybrid)	Access to production system
Vendor/integrator	IBM
Fat Node Island	

Architecture	IBM System x
CPU (vendor/type/clock speed)	Intel / Xeon Westmere-EX / 2.4 MHz
CPU cache sizes (L1, L2, L3)	L1: 32 KB per core, L2: 256 KB per core, L3: 30 MB shared
Number of nodes	205
Number of cores	8200
Number of cores per node	40
Memory size per node	256 GB
Interconnect type / topology	Infiniband QDR
Peak performance	0.078 TFlop/s
Linpack performance (measured or expected)	0.065 TFlop/s expected
I/O sub system (type and size)	Home folders provided by NetApp NAS, 1.5 PB (10 GB/s) over Infiniband Scratch and work folder provided by GPFS, 10 PB (200 GB/s) over Infiniband Backup and archiving provided by tapes plus disk cache, 30 PB, over 10 Gb Ethernet
File systems (name, type)	/home/hpc, NetApp NAS /work and /scratch, GPFS
Date available for PRACE production runs	August 2011
Link to the site's system documentation	http://www.lrz.de/services/compute/supermuc
Thin Node Island	
Architecture	IBM System x iDataPlex
CPU (vendor/type/clock speed)	Intel / Xeon Sandy Bridge-EP / 2.7 MHz
CPU cache sizes (L1, L2, L3)	L1: 32 KB per core, L2: 256 KB per core, L3: 20 MB shared among 8 cores
Number of nodes	9216
Number of cores	147,456
Number of cores per node	16
Memory size per node	32 GB
Interconnect type / topology	Infiniband FDR10
Peak performance	2.9 TFlop/s
Linpack performance (measured or expected)	2.21 TFlop/s expected
I/O sub system (type and size)	Home folders provided by NetApp NAS, 1.5 PB (10 GB/s) over Infiniband Scratch and work folder provided by GPFS, 10 PB (200 GB/s) over Infiniband Backup and archiving provided by tapes plus disk cache, 30 PB, over 10 Gb Ethernet
File systems (name, type)	/home/hpc, NetApp NAS /work and /scratch, GPFS
Date available for PRACE production runs	August 2012
Link to the site's system documentation	http://www.lrz.de/services/compute/supermuc

Table 3: Basic information of SuperMUC at LRZ

3.2 Planning of new Tier-0 systems

Two new Tier-0 systems are in the process of being purchased or deployed. The first one will be installed in Italy at CINECA and the second one in Spain at BSC. Both systems will be available to the PRACE users before the end of this year.

3.2.1 *FERMI – CINECA*

FERMI is the Tier-0 system which will be installed and managed by CINECA. It consists of a 10 rack BlueGene/Q system provided by IBM with a theoretical peak performance of 2.1 Pflop/s. FERMI installation will take place from the second half of May 2012 to the second half of August 2012. The system is expected to be in full production (i.e. opened to users) by September 1st, 2012.

FERMI compute and I/O nodes are managed by a fully featured RHEL 6.X based Linux distribution. A 5-dimensional torus network interconnects all compute nodes with an embedded collective and a global barrier network. The I/O nodes connection to the Storage Server Cluster will be realised with an InfiniBand networks.

The connection to the CINECA shared storage repository and archiving facility is performed via a 10Gb Ethernet.

The system will be provided with 8 Linux-based frontend nodes. Four of these frontend nodes will be used as login nodes where users will perform interactive access and job submission operations via the IBM Tivoli Workload Scheduler LoadLeveler. The remaining four frontend nodes will be used as data service nodes to perform either data transfer or archiving operations. At least two frontend nodes will be directly connected to the PRACE Network infrastructure.

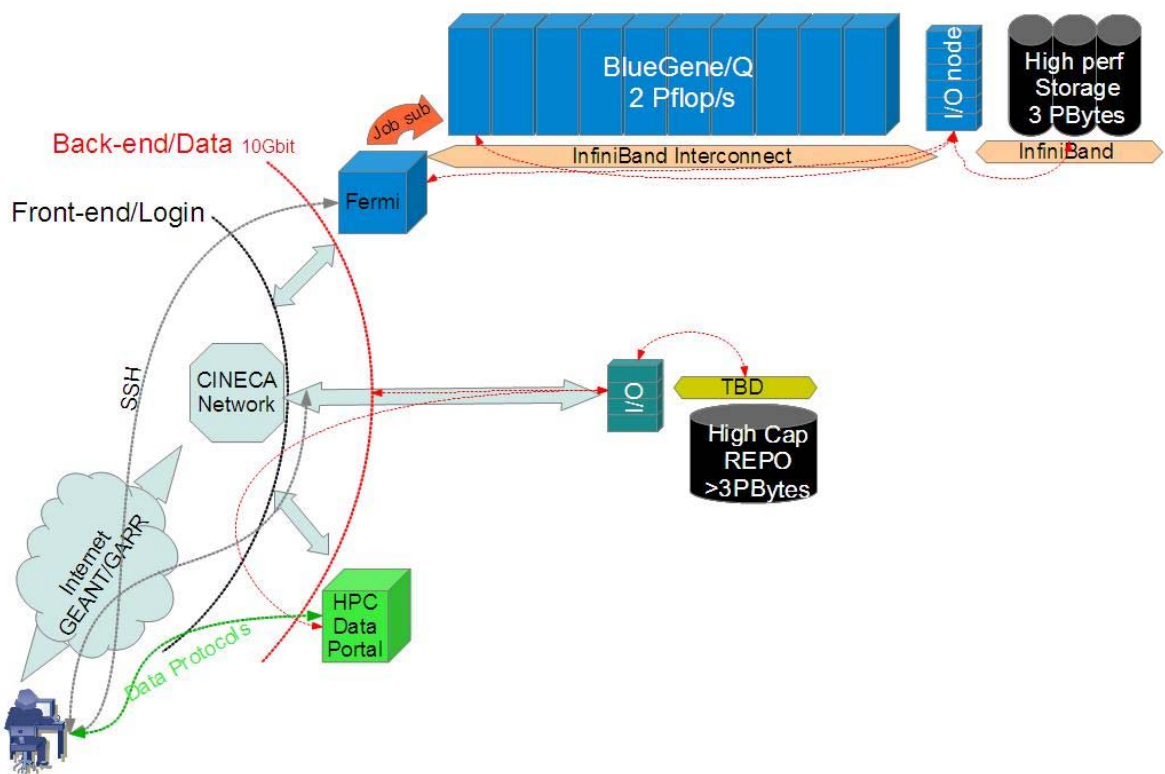


Figure 10: FERMI network and storage schema

IBM GPFS (General Parallel Filesystem) technology will provide access to both the local storage area (home, scratch) and the global storage area (shared data repository). Access to the central archiving facilities will be granted by the IBM Tivoli Storage Manager.

Machine name	FERMI
PRACE partner	CINECA
Country	Italy
Organisation	CINECA
Location	CINECA, Casalecchio di Reno, Italy
Nature (dedicated system, access to system, hybrid)	Access to production system
Vendor/integrator	IBM
Architecture	Blue Gene/Q
CPU (vendor/type/clock speed)	IBM / PPC A2 / 1.6 GHz
CPU cache sizes (L1, L2, L3)	32 MB shared
Number of nodes	10240
Number of cores	163840
Number of cores per node	16(compute)+1 (control)+1(spare)
Memory size per node	16 GB
Interconnect type / topology	Proprietary / 5D- torus + tree
Peak performance	2.1 Pflop/s
Linpack performance (measured or expected)	1.7 PFlop/s expected
I/O sub system (type and size)	InfiniBand connected GPFS server Approx. 3.6 PB raw disk space

File systems (name, type)	/fermi/ (HOME directory), /gpfs/scratch/ (SCRATCH filesystem), /shared/data (Data repository) (all GPFS)
Date available for PRACE production runs	September 2012
Link to the site's system documentation	http://www.cineca.it/en/hardware/ibm-bgq

Table 4: Basic information of FERMI at CINECA

3.2.2 *MareNostrum* – BSC

MareNostrum, the new Tier-0 hosted by BSC in Barcelona, Spain, will be announced shortly and the configuration of the system cannot be disclosed at the time of this writing. The system deployment will follow a stepwise approach; the first deployment phase is planned to be completed by the end of 2012.

4 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.

Common services are divided into thematic categories: Network, Data, Compute, AAA, User and Monitoring. Each service category has a responsible person who is in charge of managing all the information and decisions related to a specific service area.

Selection of common services is ruled by the PRACE service catalogue and once chosen, each service is then taken in charge by the respective service area.

Intensive use of the PRACE Wiki has been made since its deployment at the beginning 2011. This wiki is the central collaborative tool used to coordinate all deployment, test and other operational activities undertaken for PRACE common services.

The following sections provide the current status of each service category and the main steps achieved within the past year.

4.1 Network services

The PRACE network services are based on the developments done in the DEISA project. Any network services are delivered within PRACE-2IP. Operational work within PRACE-1IP in the period after the end of DEISA and before beginning of PRACE-2IP is delivered as in kind contribution.”

4.2 Data services

GridFTP is a data transfer protocol that can fully utilize the high bandwidths between the PRACE Computing centres, so it has been picked as standard for the sites.

GridFTP supports parallel TCP streams and multi-node transfers to achieve a high data rate via high bandwidth connections. Furthermore, transfers can be restarted and third-party transfers can be established, what is very useful to the PRACE users.

4.2.1 *Status of Deployment*

In May 2012, the following Tier-0 sites have deployed GridFTP Services for their systems:

Site	Version
CEA	3.28/GT5.0.3
FZJ	3.23
HLRS	3.33/GT5.0.4
LRZ	3.33/GT5.0.4

Table 5: Deployment Status of Tier-0 sites

In addition, several Tier-1 sites have installed the service, too. The deployment status is monitored at the inventory page in internal the PRACE Wiki.

4.2.2 Documentation

The internal documentation how to install and setup the GridFTP service at a PRACE site has been maintained and further tested for new versions of the GridFTP Server (3.28, 3.33, 6.5) resp. the Globus Toolkit (5.0.3, 5.0.5, 5.2.0).

In addition, user documentation has been developed and is now provided on the official PRACE Website [1]. This documentation describes the process of retrieving data from a users perspective and explains the syntax of the command-line tool *globus-url-copy*, the standard tool delivered within the Globus Online Toolkit.

4.2.3 Secure setup

At HLRS, deployment of this service has been finished in March 2012 with a specific setup due to security restrictions. This setup is also emphasized in internal the PRACE Wiki.

The service internally consists of two parts, the frontend for negotiation of the transfer-process and the backend for user-authentication, access to the supercomputers file system and the transfer itself. In the secure setup, these two parts are separated to two distinct machines, where the users only see the frontend node. To the users, it seems like there is only one machine involved. If this exposed machine gets compromised, the attacker cannot get access to the file systems of the supercomputers, because they are only accessed by the backend.

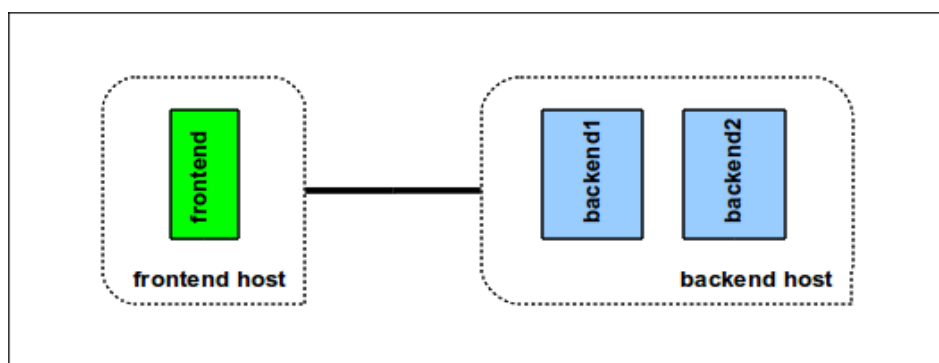


Figure 11: Principle Setup of the Tier-0 GridFTP at HLRS

This setup is depicted above. Details on the GridFTP protocol are provided PRACE internally for the site administrators.

4.2.4 Advanced User Tools

Progress regarding the *gtransfer* tool can be found in the chapter for Task 6.3.

4.2.5 Monitoring

Monitoring of data services takes place using the PRACE Inca Monitoring. Tests are performed from each GridFTP server to another, so the mutual availability is assured. As an example, the reachability from SARA to the other sites at an arbitrarily chosen point in time is shown below.

PRACE GridFTP

Byte Transfer	SARA Huygens
gridftp to bsc	pass
gridftp to cea-curie	error
gridftp to cea-inti	error
gridftp to csc	pass
gridftp to epcc-hector	error
gridftp to fzj	pass
gridftp to hlr	pass
gridftp to idris	pass
gridftp to lrz-super	pass
gridftp to rzg	pass
gridftp to sara	pass

Figure 12: Monitoring of GridFTP status

4.3 Compute services

The activities carried out in the compute services area for Tier-0 systems, during the second year, focused on the consolidation of the documentation (installation and configuration guides), deployment monitoring, support and information gathering.

Two types of interaction are used between users and compute services:

1. Direct interaction with a local job management framework provided by each computing system;
2. A seamless access to all distributed resources through a common software layer built on top of the computing infrastructure.

Future and alternative options can be provided by new computing paradigms like HPC Cloud Computing, which is considered today a promising technology but with still unsolved issues in security and performance.

The processes of selection, deployment and configuration vary a lot between the two functionalities listed above.

4.3.1 Local Batch Systems

For direct interaction, all sites are obviously free to select, deploy and configure the best combination of Resource Manager and Scheduler that fits with their specific hardware platform, software environment and other specific requirements, like the way to implement the accounting management for example. What has been done here was to create an inventory for collecting all information related to each site implementation. It is named the “Batch

System Inventory” and it is available in the PRACE Wiki since the beginning of the PRACE-1IP. This second year it has been enhanced with the information about new Tier-0 systems (SuperMUC@LRZ and Hermit@HLRS), see also Table 6.

Site	System	BSS	Arch	MicroArch	Nodes	Cores
FZJ	JUGENE	LoadLeveler	PowerPC	Power PC 450	72 728	294 912
CEA ⁽¹⁾	CURIE	SLURM	x86-64	Nehalem-EX (X7560)	360	11 520
				SandyBridge-EP (E5-2680)	5 040	80 640
HLRS	Hermit	Torque/Moab	x86-64	Interlagos (6276)	3 552	113 664
LRZ ⁽²⁾	SuperMUC	LoadLeveler	x86-64	SandyBridge-EP	9 216	147 456
				Westmere-EX	205	8 200
(1): 1 st row is for Fat Nodes, 2 nd for Thin Nodes. GPU cores excluded.						
(2): 1 st row is for Thin Nodes, 2 nd row for Fat Nodes.						

Table 6: Batch System Inventory for Tier-0 Systems

Even though the data only covers four systems, the situation seems to reflect the typical heterogeneity that usually characterise the market of batch systems, with LoadLeveler as one of the market leaders and SLURM, a well-known open source solution capable of handling the allocation of resources made by a large pool of computing nodes.

Due to the autonomy of each centre, a common approach for batch systems documentation is required in order to know what kind of features are provided to users and how to document them in a common format.

4.3.2 UNICORE

The second way to interact with compute services is through a unified layer built on top of the Batch Systems. To implement this type of interaction, the deployment of UNICORE [14] has been selected for Tier-0 systems.

UNICORE comes with different software components, which are responsible for the entire orchestration and management of the execution of a job. Through UNICORE a user can define a job in a seamless way without concerning about the underlying batch system. The added value provided by this abstraction layer is more evident in a Grid environment, where users usually move from one system to another. In the case of the Tier-0 infrastructure, benefits are, however, evident in terms of easy management of the executions. **Table 7** summarises the portfolio of software components included in the deployment process for Tier-0 systems.

Component	Type	Description	Deployment Site
Registry	Server	The Registry is a directory service contacted by the clients in order to connect to the Tier-0 network	FZJ (Primary site), CINECA (Backup site)
Unicore/X	Server	The UNICORE/X component is the central server which translates the abstract jobs into concrete jobs for the target system, submits and monitors the jobs	Installed at each Target Tier-0 System

Gateway	Server	The main entrance to each Tier-0 system, through which the internal components can be reached. All client connections will be with the gateway, which will then forward the requests, and send the replies back to the client.	Installed at each Target Tier-0 System
XUADB	Server	The XUADB is responsible for user authentication and authorization, it is used to map user credentials (such as an X.509 certificate or X.500 distinguished name) to a set of attributes and it is synchronised (manually or automatically) with the PRACE LDAP entries.	Installed at each Target Tier-0 System
TSI	Server	The Perl (also called legacy) TSI component communicates with the local systems (batch scheduler or file systems)	Installed at each Target Tier-0 System
Unicore Rich Client (URC)	Client	The URC is a graphical client suite based on the Eclipse framework and offers a rather complete set of job management operations.	User's home institution
Unicore Commandline Client (UCC)	Client	UCC is a commandline and extensible client offering a set of basic operations (run job, get output, transfer files, etc).	User's home institution

Table 7: UNICORE software components deployed on Tier-0

The deployment design that has been implemented within the Tier-0 infrastructure is shown in Figure 13.

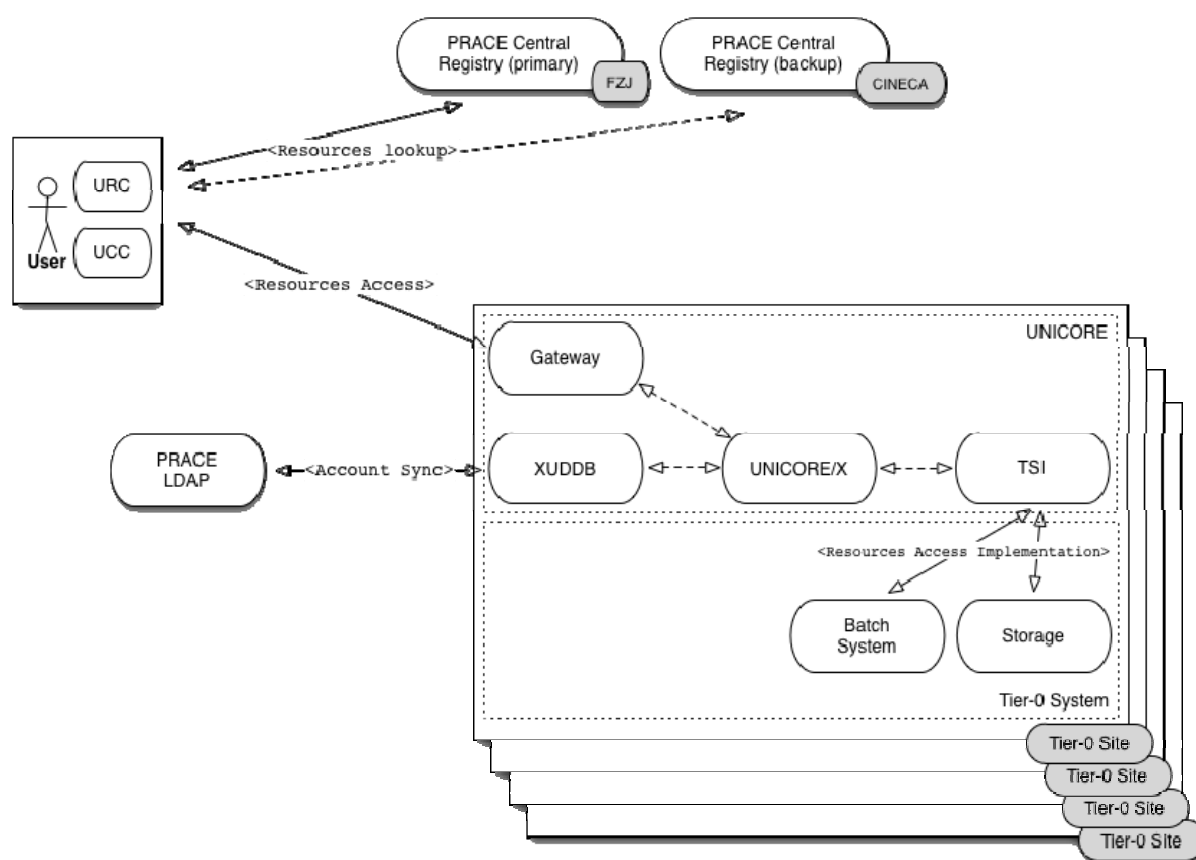


Figure 13: UNICORE deployment design on Tier-0 systems

D6.3 Second Annual Report on the Technical Operation and Evolution

The actual default version is the 6.4.2-p2, released on February 2012. Before triggering the deployment process, this release has been tested even though it is considered a minor change of the previous used version (6.3). The pre-production test has been carried out and are defined as follows:

- Testbed; 3 systems have been considered:
 - o Jugene@FZJ (Tier-0);
 - o MareNostrum@BSC (Tier-1);
 - o JuRoPa@FZJ (Tier-1).
- Test Cases; 3 test cases have been considered
 - o Test.A: Submission of simple job script
 - o Test.B: Submission of complex job script including file stage in and stage out
 - o Test.C: Stress test of server components
- Users; 4 users belonging to PRACE Staff have joined the test by connecting from 4 different sites (BSC, CEA, FZJ, SARA).

The test was successful and the deployment started in March 2012. Activities are currently monitored through the wiki and updated by site partners. The actual status is shown in **Figure 14**, which is a snapshot taken by the PRACE Wiki itself.

Central services are available with the new version. Unicore 6.4.2-p2 is installed on 2 Tier-0 systems, a third Tier-0 is running a “not fully updated” version (6.4.1) which is totally compatible with the current one while only one Tier-0 system has not yet completed the deployment process. In order to avoid any confusion with the picture, the “Workflow Engine” component is considered as optional, that is Tier-0 sites can optionally provide it while availability and long-term support are not guaranteed by PRACE.

Deployment of Central Services

Central Registry	Hosted by	URI	Status
Primary	FZJ	https://prace-unic.fz-juelich.de:9111/PRACE/services/Registry?res=default_registry	Upgraded to 6.4.2 [OK]
Backup	CINECA	https://grid.cineca.it:9111/Prace/services/Registry?res=default_registry	Upgraded to 6.4.2 [OK]

Deployment on T0 systems

Site	System	Network	Version	U6 gateway (url)	U6 unicorex (url)	U6 xuudb	U6 TSI	Workflow Engine?
CEA	CURIE	public	6.4.2-p2	gruget.eole.ccc.cea.fr	burle.eole.ccc.cea.fr	installed	installed	no
FZJ	JUGENE	public	6.4.2-p2	prace-unic.fz-juelich.de	prace-njs.fz-juelich.de	installed	installed	yes
HLRS	Hermit	public	(no info)
LRZ	SuperMUC	public	6.4.1	uc6.lrz.de	ucx.lrz.de	installed	installed	no

Figure 14: UNICORE deployment status on Tier-0 systems

The software repository for UNICORE is hosted by SourceForge[15].

4.4 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.

4.4.1 *Public Key Infrastructure - PKI*

Several PRACE services rely on X.509 certificates [6] for the authentication and the authorization. These certificates must be issued by entities which are trusted by the service providers. 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, as published in a profile document.

For PRACE a distribution of CA information is maintained at a central repository [7]. The distribution is provided in several formats because services have different requirements for the presentation of the information. In this project period eight new IGTF distributions have been made available for the PRACE infrastructure.

4.4.2 *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 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.

A single LDAP server is used for PRACE Tier-0 accounts. The LDAP infrastructure for Tier-0 and Tier-1 accounts is tightly integrated, which is shown in **Figure 15**. The top part, the suffix, for PRACE Tier-0 accounts is “ou=ua,dc=prace-project,dc=eu”, while for Tier-1 accounts this is “ou=ua,dc=deisa,dc=org”. This means that two databases are used, however for account information the same LDAP schemas are used, so the same tools can be used to update and retrieve information. The difference in suffix only has an historical reason and doesn't have any functional reason. The change to one suffix is planned.

Three LDAP domains (branches) for Tier-0 accounts exist: FZJ, CEA and HLRS. Each of these partners manages the Tier-0 accounts for their Tier-0 system: JUGENE (FZJ), CURIE (CEA) and HERMIT (HLRS). Additional branches will be created for the other three Tier-0 systems which are planned to become operational later in 2012.

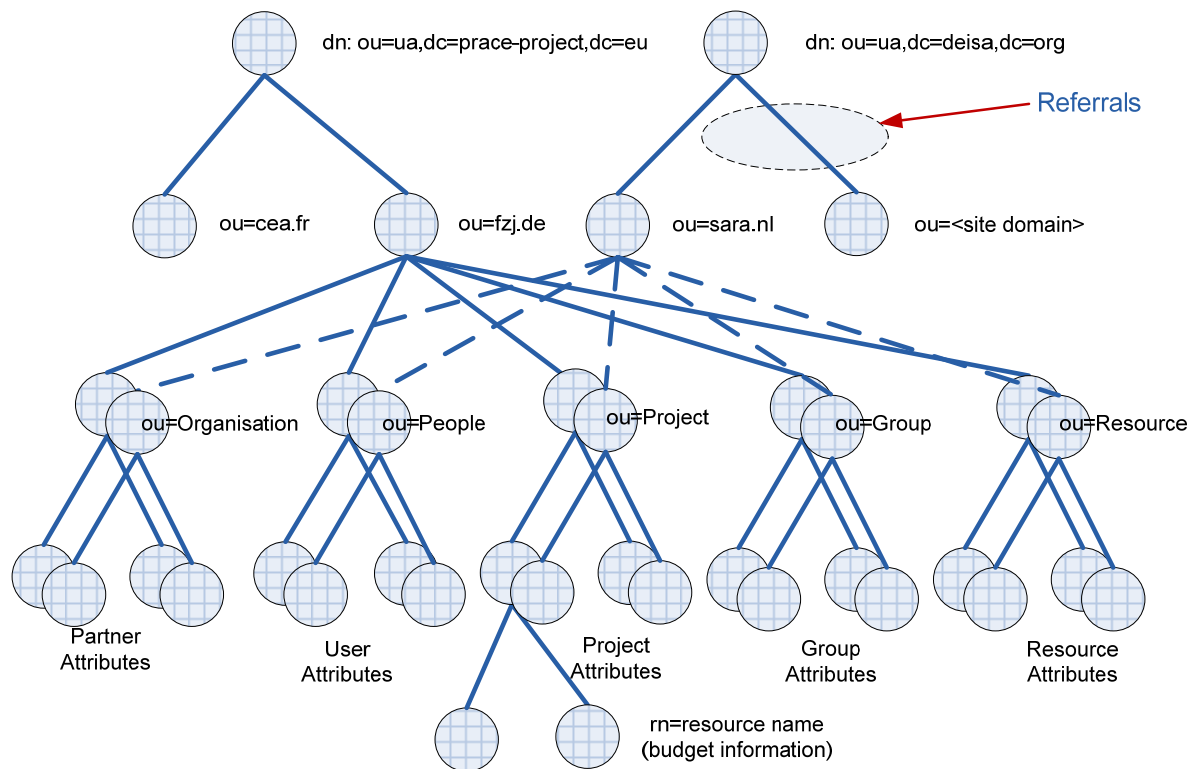


Figure 15: PRACE LDAP directory tree

The main service is operated by SARA and for high availability a local replica server is operational, too. Since early 2012 a backup server is operational at HLRS, too. This server contains all data from all PRACE LDAP servers and can be used in case the production servers at SARA cannot be reached.

The PRACE AAA Administration guide has been updated with a section for the user administration. This describes the general set-up, the security requirements and the policies and procedures for the service.

The AAA team discussed some open issues, also for the Tier-1 account administration, and several changes have been implemented.

4.4.3 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. For interactive access between PRACE sites and for some sites also from external the use of X.509 certificates for authentication is preferred. The Globus community distributes a X.509 based OpenSSH version, GSI-OpenSSH [8] or GSI-SSH for short. On JUGENE and CURIE GSI-SSH based access is enabled. GSI-SSH_Term, a GSI-SSH client, is supported by the PRACE partner LRZ.

4.4.4 Accounting services

Information about the usage of resources is important for users, Principal Investigators (PIs), partner staff and the management of the resources. PRACE provides facilities to publish and display usage with the following characteristics: 1) the usage of resources is published in a common format, which follows the recommendations of OGF's UR-WG (Usage Record Working Group) [9]; 2) access is based on the authorizations of the requestor, e.g. a normal

user can only see his/her personal usage while the principal investigator of a project can see all the usage of his project. Detailed information about the design considerations can be found in [10].

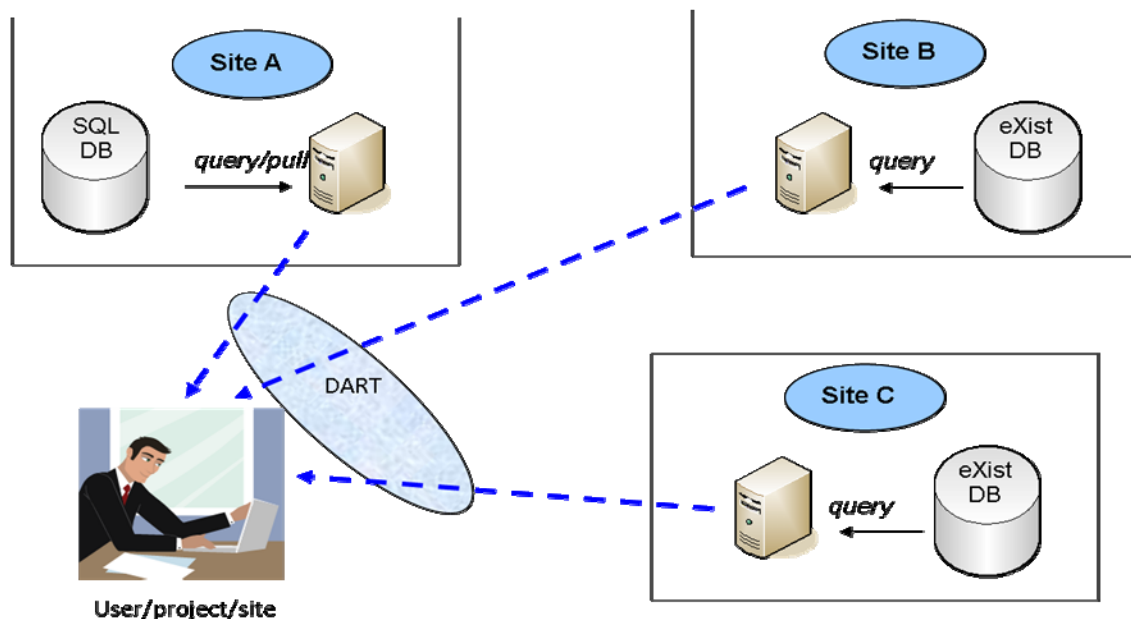


Figure 16: Accounting architecture

Figure 16 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). An Apache/CGI web interface is available which will provide data to authorized clients. The authorization is based on X.509 certificates and the access rights are given by the attribute `deisaAccountRole` of the user administration service. DART [11] is a Java Webstart tool, which can be used by a client to retrieve and to display the information from different sites.

For JUGENE at FZJ the facilities are implemented. For CURIE at CEA usage information is available locally; external access is waiting on the implementation of the authorization facilities.

4.5 User services

4.5.1 PRACE Common Production Environment

The PRACE Common Production Environment (PCPE) distribution has been updated to allow for installation on Tier-1 sites as well as Tier-0. Due to the larger range of architectures present at the Tier-1 level compared to Tier-0 this has entailed introducing more flexibility for individual sites on deciding how the components of the PCPE are implemented locally. In particular, the restrictive nature of the naming and organisation of PCPE modules was relaxed so that sites can take advantage of existing module installations which are becoming more and more commonplace in HPC installations. This increased flexibility has also solved the issue that the DEISA CPE (DCPE) could not be installed on Cray systems; the PCPE is now deployed successfully on almost all Cray systems in PRACE.

The PCPE assumes that sites have the modules software available (although it can be implemented purely using shell scripts as a temporary measure if needed). The PCPE guarantees that a certain minimum set of software tools and libraries are available on each

PRACE site. The actual module names for each of the components of the PCPE can be decided by each site. One module called 'prace' is defined that loads all the required modules and sets the PRACE environment variables. The command: `module load prace` is required at all sites to load the PCPE. Site administrators can also add additional tools/libraries to the PCPE at their site if they think they are beneficial to users.

The PCPE is set up to ensure that if a user is compiling, all they need to do is add the environment variables `$PRACE_FFLAGS` (Fortran) or `$PRACE_CFLAGS` (C, C++) to their compile line to get all the correct include options and compile flags for the tools/libraries that are part of the PCPE. Similarly, at the link stage users would just need to add `$PRACE_LDFLAGS` to their link line to link to PCPE guaranteed libraries.

Use of the PCPE has been documented in the PRACE User Documentation.

Currently, The PCPE has been deployed on the Tier-0 systems at FZJ, HLRS and CEA. The PCPE has also been deployed on the majority of Tier-1 sites that are currently active in the DECI calls.

We are currently working to integrate the monitoring of the PCPE status into the PRACE Inca monitoring tool.

4.5.2 User Documentation

User documentation for PRACE is now available online on the PRACE website at <http://www.prace-ri.eu/Documentation-and-User-Support> [1]. The PRACE documentation is split into a number of subcategories. To date ten different documents have been published covering core topics such as User FAQs, Compute, Batch Systems, Data Transfer, PRACE Helpdesk and others.

Documentation version control is done via the central PRACE SVN repository. Document owners have been defined for all areas. Edits are performed by the document owners in the PRACE SVN, with the ability to upload the documents to the PRACE Website being restricted to the Documentation task leader and a deputy. The document owners are all part of the Documentation Review Panel. Any major changes or new documents must be reviewed by the panel before they will be posted online. Minor changes can be routed directly to the Documentation lead for publication.

There is a requirement to create paper documentation which can be downloaded as well as online documents. Work is in progress to facilitate this. This is covered in Chapter 5.

4.5.3 The PRACE Trouble Ticket System

The centralised PRACE Helpdesk is now in operation servicing all Tier-0 facilities.

The primary interface to the Helpdesk is via a simple web interface. The web interface allows PRACE users to submit new queries and to monitor any existing queries that they have in the Helpdesk system. Authentication to the web interface is based on having an X.509 certificate imported into the user's web browser.

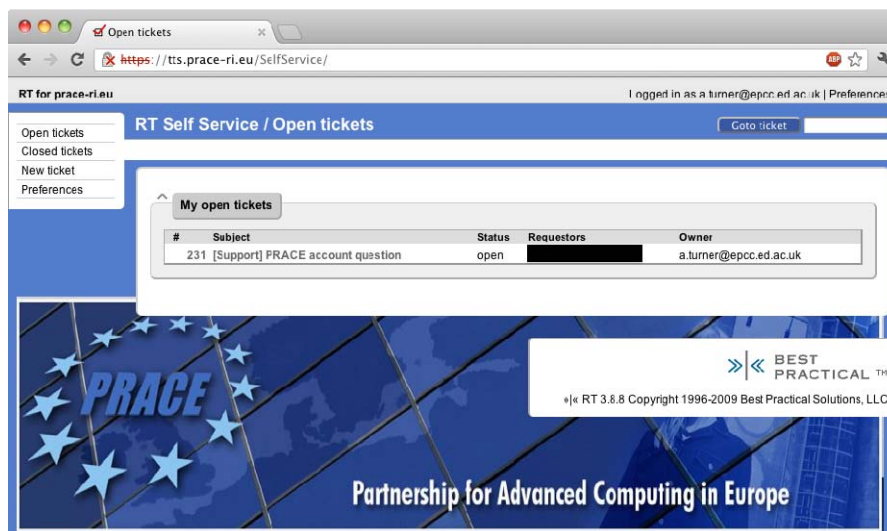


Figure 17: PRACE TTS Self Service Interface

Issues raised via the web interface are routed automatically to the Contributor, thus minimising any delay in receiving support.

A secondary email-based interface is also now in operation. This option is required if a user is 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) has been configured to route to the PRACE TTS. Issues raised in this manner are monitored by the duty Helpdesk team, who in turn route the issue to the appropriate Contributor.

Furthermore, for Tier-0 sites direct email addresses have been configured. The email addresses route the user request to the Helpdesk, but rather than route them to a general queue which needs operator intervention, they route directly to the support team for the Tier-0 site. These email addresses take the format <site-resource>-support@prace-ri.eu e.g. cea-curie-support@prace-ri.eu

Full user documentation for the Helpdesk is available on the PRACE Website at <http://www.prace-ri.eu/Helpdesk-Guide>.

Support staffing effort for the Helpdesk is provided by the PRACE partners on a rotational basis, with each site manning the Helpdesk for one week at a time. The Helpdesk on Duty role is to ensure that any 'General' queries are routed appropriately, are correctly categorised, and that all open requests are being resolved by the Contributors in a timely manner. The categorisation of tickets is important to ensure that incidents can be measured and that associated KPIs can be defined and managed. A wiki based handover mechanism is also in place to ensure that any open issues can be passed consistently from one partner to the next.

4.6 Monitoring services

The monitoring subtask implements, deploys and operates tools for monitoring of PRACE e-Infrastructure services. The subtask focuses on availability and functionality monitoring of services from the e-Infrastructure and user perspectives. Among all monitoring solutions available on the market only a few are well suited for service monitoring from the user perspective. PRACE selected a user-level monitoring application called Inca [12]. The decision was primarily based on the positive experience and recommendations received from

the DEISA project that used *Inca* for user-level monitoring for over four years. More information about *Inca* and its usage in DEISA can be found in PRACE deliverable D6.2 [3].

Inca implements the client-server model, where *Inca* reporter managers clients that are deployed on all PRACE resources are testing components of the PRACE e-Infrastructure. Collected monitoring data is then sent to an *Inca* server for processing, archival and presentation. In PRACE the latest *Inca* version 2.6 is deployed. *Inca* server components are installed and running at LRZ in a virtualised environment that guarantees efficient load balancing and high fault tolerance. *Inca* reporter managers are operational on 15 resources among which 13 are PRACE HPC systems and 2 resources used for monitoring of PRACE central services. Integration of 8 further systems, including HLRS Hermit Tier-0 systems that recently went into production and PRACE Tier-1 systems contributed by new partners, is ongoing.

Inca reporters that test specific service functionality are developed and configured in accordance with service requirements outlined in the PRACE Service Catalogue (see Appendix A). Currently Common Production Environment components, Globus GSISSH, Globus GridFTP, LDAP User Administration and UNICORE services are being monitored. *Inca* tests are grouped together in suites based on an individual subtask domain to facilitate problem reporting and resolution. Services that are deployed on all or a majority of PRACE resources and are accessible from any PRACE system, for instance Globus GridFTP, are monitored in all-to-all fashion. For instance, in case of GridFTP, data transfer tests are performed between all possible clients and all GridFTP service endpoints available on the PRACE backbone network. Such testing is necessary to ensure availability of PRACE services to all users and guarantees correct functionality of services within the e-Infrastructure. Additional *Inca* reporters will be developed throughout the course of the project to satisfy e-Infrastructure operation and user requirements.

Inca enhancements were migrated from *Inca* deployed in DEISA or developed in the scope of the PRACE project to streamline daily operational tasks. This includes an interface to PRACE Wiki for accessing resource and service maintenance information and an interface to PRACE Trouble Ticket system for facilitating trouble ticket creation and management. The PRACE Wiki interface allows *Inca* to display resource and service maintenance information and helps PRACE operation staff to filter monitoring results based on the expected service availability. By interfacing PRACE Trouble Ticket system *Inca* is able to assist PRACE Operator on Duty in reporting on detected incidents and to link exiting tickets that describe detected infrastructure problems. References to the Trouble Ticket system functionality are available across multiple *Inca* monitoring result views and can apply to individual as well as series of tests, for example equivalent tests failing on several resources or test suites that fail on a single resource.

PRACE *Inca* configuration and functionality is documented in detail with all documentation available to PRACE staff members. *Inca* installation and administration guides as well as a detailed description of relevant features and suite configuration are maintained in PRACE Wiki. *Inca* reporters and supplementary tools, such as reporter manager initialisation script, are stored in PRACE repository.

As mentioned above, *Inca* server components are deployed in a virtualised environment that guarantees efficient load balancing and high fault tolerance. However, in exceptional situations, such as power failure, *Inca* server components might become unavailable. To avoid complications that might be caused by such incidents it was decided to deploy and operate a back up *Inca* instance. *Inca* back up will be set up at FZJ over the course of next few months. *Inca* depot will be deployed in a mirror mode so that depot instance running at FZJ will be continuously synchronised with *Inca* depot running at LRZ. This will provide additional fault

D6.3 Second Annual Report on the Technical Operation and Evolution

tolerance and facilitate failover to the backup Inca instance in case of problems with the primary installation.

5 Identification, selection and evaluation of new technologies

This chapter presents the results of the work that has been performed within the Task 6.3 during the second year of the project for the identification, selection and evaluation of new technologies. Besides this, the task has also worked to improve the quality of existing services, to strengthen the collaboration with other EU projects (see paragraph 2.5), to develop a data management strategy and to design new services in order to meet unaddressed user's requirements. Major achievements are reported in the following sections as well as the activities that will be carried on in the course of the PRACE-2IP project. The establishment of a collaboration with PRACE-2IP WP10 was one of task objectives as it is fundamental to guarantee a synergy between the two projects in the evolution of the PRACE Infrastructure towards a common target.

Thus the major part of allocated effort has been devoted to carry on the following activities:

- consolidate and extend existing services evaluating new technologies on the base of collected requirements (see par. 5.4.1, 5.4.2, 5.4.3, 5.4.4, 5.4.5, 5.4.6);
- enforce the collaboration with other EU projects (i.e. MAPPER, EMI, IGE etc.) for improving the sustainability of adopted technologies (see paragraph 2.5);
- improve the quality level of offered services through the definition of a service certification process (see par. 5.2);
- prepare a proposal for developing a new service (see par. 5.3) to improve user's experience while interacting with the PRACE Infrastructure;
- extend the accounting systems to provide more information about allocated resource budget (see par. 5.4.4);
- lay the foundation for developing a data management strategy (see par. 5.4.2).

The remaining part of the chapter is organized as follows:

- Paragraph 5.1 reports the results of the activity concerned with the gathering of user's requirements;
- Paragraph 5.2 presents the process for the certification of PRACE services;
- Paragraph 5.3 describes the proposal for the development of a PRACE Information System where to collect and present information about the status of the PRACE Infrastructure;
- Paragraph 5.4 gives an overview of the outcomes within the different service areas;
- Paragraph 5.5 is the summary of most relevant achievements for Task 6.3.

5.1 Requirement analysis

On November 2011, a new user survey was prepared with the intention to better understand users' experience and requirements of the PRACE Tier-0 systems usage. The survey consisted of a set of 48 questions that were to be answered by the users of PRACE Tier-0 systems being involved in the PRACE Access projects, including both Preparatory Access and Regular Access projects. The responses were collected between December, 13th, 2011 and March, 5th, 2012. A total of 62 valid responses were received. A first review of collected responses was conducted with the co-operation of WP7, Task 4 (Applications requirements for Tier-0 systems) and was mainly concerned with the analysis of application's requirements. A second review will take place during the course of the PRACE-2IP project when specific requirements on the service layer will be further analysed.

The surveys were organized as follow:

- a survey of the PRACE Tier-0 systems, JUGENE and CURIE;
- a survey of the PRACE Tier-0 Access users, including both Preparatory Access and Regular Access.

The purposes of these surveys were:

- to understand the current usage status of the PRACE Tier 0 systems;
- to understand the users' experience and further requirements for the PRACE Tier-0 systems usage.

Preliminary outcomes of the first review can be summarised in the following list:

- parallelisation method mostly used is: MPI, OpenMP & MPI;
- programming languages: Fortran, C;
- memory size per core required for your typical production jobs : less then 0,5 GB;
- architecture features desired the most: “Higher peak flop rate” scored the highest, followed by “Lower point to point communications latency”;
- GPU support: Around 61% of the responses indicated the applications have accelerator implementations or may potentially benefit from accelerators;
- PRACE dedicated network – not used explicitly by most users;
- LRMS submission prevails UNICORE in the significant percentage;
- systems are accessed directly in most cases;
- no perceived benefits for X.509 usage for most users;
- accounting info is important;
- visualisation services deserve an important role. A dedicated task is covering this topic within the PRACE-2IP project.

The full overview of the survey results is presented in deliverable [20].

5.2 Service certification

A service certification procedure has been defined within PRACE in order to manage the quality of services delivered to the users by collecting information from existing frameworks and procedures such as monitoring, and by performing additional custom quality checks. 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 the PRACE’s offering meets user’s expectations. Thus the main goals of Service Certification procedure can be summarized as follows:

- **verify:** deployed services before offering them to the users;
- **ensure:** that expected functionalities are supported;
- **ensure:** that technical requirements (e.g. non-functional requirements) are implemented;
- **control:** that quality standards, such as operational policy are satisfied;
- **improve:** overall quality of offered services;
- **provide:** probes complementary to live monitoring.

This section presents a proposal for the correct certification of services within PRACE, including a possible certification process, the list of quality characteristics to be measured, a sample check list with the controls to be performed on a specific service.

When to perform certification

The certification of a new service could be performed:

- before the service enters the production level (*Acceptance test*)
- after any major change to the configuration on the hosting system (*Validation test*)
- once in a while (i.e. every six months, every year, etc.) (*Validation test*)

There should be also a possibility of revoking the certification status for a service instance in case it fails to deliver required quality of service for long period of time.

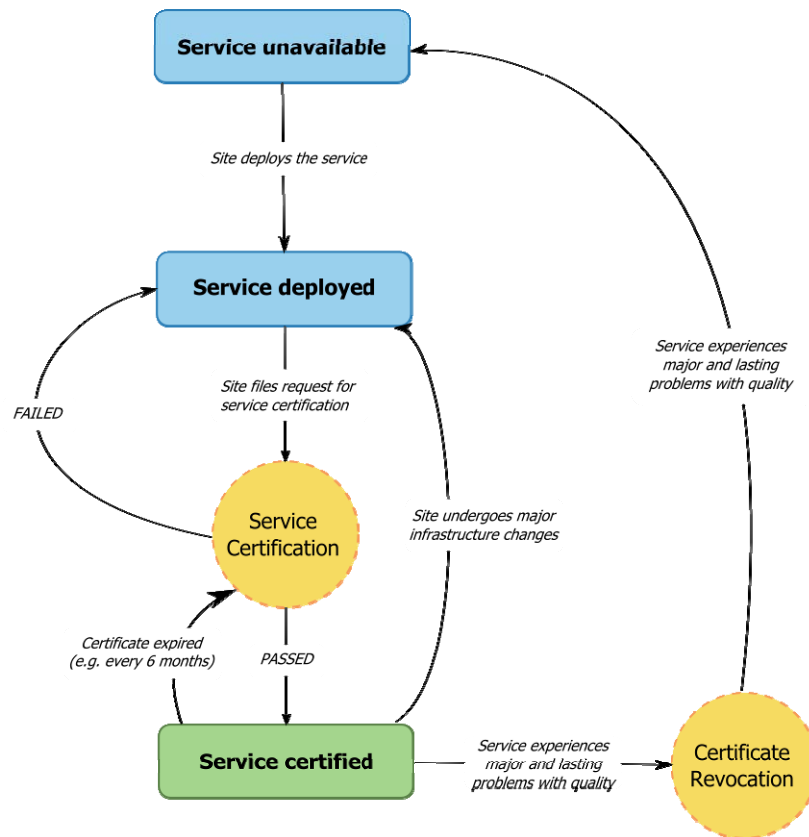


Figure 18: Service Certification state diagram

The generic certification process for a service is presented in Figure 18. At first the service is not available on a given site. After deployment the site reports that the service is deployed and configured, and requests the relative Service area Leader to certify it. If the certification phase fails, the service does not get a certified status and some feedback is given to the site on what to improve in order to pass next certification process. In case the service passes relevant quality checks, the service gets the certified status and enters the PRACE services portfolio. Only the services that have passed the certification control are offered to the users. The certified status can change due to several factors. First of all, each certificate is valid only for a given period of time (e.g. 1 year) although this period of time might depend on the service's nature. Furthermore, every time the site undergoes major infrastructure changes (the bottom line is that the service had to be redeployed because of the change) the service should be automatically filed for certification in order to ensure that the changes did not have any impact on the service quality. Finally, when the service fails to meet certain Quality of

Service parameters (such as KPIs) for a longer period of time, the certificate revocation process is initiated which can involve an intermediation phase with the site in order to provide them feedback on necessary changes. After fixing the problem and redeploying the service the site can request a new certification.

Quality checklists

This section contains quality control checklists for the core PRACE services listed in Appendix C for the purpose of implementing the service certification procedures. The quality control checklists will allow verification whether deployed services meet certain quality (and functionality) standards as defined in the requirements and service catalogue documents.

Services in PRACE are classified along two main groups of categories. First of all, services can belong to one of these classes: a) Core, b) Additional, c) Optional. Secondly, services are grouped into six service areas: a) Network, b) Compute, c) Data, e) AAA, f) Monitoring, g) User.

Each quality control list contains several “check points” which have to be verified manually or, if possible, automatically using special scripts (we recommend to create a special script based framework for automation of validation of most of the check points). Each quality check list is organized into categories based on the characteristic of the quality check. Relevant quality characteristics include:

- Accessibility – these checks verify whether the service is available under a declared access point from various locations depending on the requirements;
- Compatibility – the service is compatible with the standards and procedures of other PRACE services;
- Documentation – these steps must verify that each service should provide Users and Administrators Guide as well as brief Service Reference Card, and that the documentation is of good quality;
- Security – these checks must ensure that the service is securely deployed and configured, conforming to the PRACE Security Forum rules and standards;
- Stress – these checks must verify that the service can handle certain peak load situations (e.g. number of users, connections, transfers);
- Usability/Functionality – these checks verify that the basic functional requirements of the service (as defined in the PRACE Service Catalogue) are working;
- Administration – these checks verify that the administration of the service is on certain level (e.g. that there is a clear administrative responsibility, logs are kept for certain period of time, etc.).

Each quality check point consists of a unique ID, textual description of the quality check, severity level of the requirement (HIGH, LOW) and finally the check status (‘+’ - passed, ‘-’ - failed, ‘?’ – problems with verification).

Depending on the type of the checkpoint several techniques can be applied such as:

- Document review
- Source code review
- Standards compliance validation
- Scenario testing
- Prototyping
- Simulation

- Walkthrough
- Automated scripts

First version of the quality checklists was defined for a selected subset of services from the latest version of the PRACE Service Catalogue and these include:

- Uniform access to HPC
- PRACE internal interactive command-line access to HPC
- Data transfer, storage and sharing
- Authentication
- Authorization
- Accounting
- Network management
- Monitoring
- Software Management and Common Production Environment

The current version of the quality checklists can be found in the PRACE wiki. Example of a quality checklist (for Uniform access to HPC) is attached in Appendix C.

Implementation

The implementation of the Service Certification procedures and quality checklists is planned for the PRACE-2IP, however some basic requirements and vision have been sketched within WP6 T6.3 of PRACE-1IP.

The overall idea is to implement an automated web based framework, which will support services certification process. For each service, it will follow a particular scenario of the general form:

- take all necessary input from the user performing the certification (site, machine, temporary certificates, test files, etc.);
- perform automatic tests (run specified scripts);
- prompt the user to perform manual checks (e.g. document review, user interface, etc.);
- collect user input;
- store the certification status.

Due to the nature of the service certification, the entire process cannot be fully automated thus the implementation might involve a manual intervention from Operational Staff. A tentative vision of the architecture of the PRACE Service Certification platform is presented in Figure 19. The basic idea is to reuse as much information as possible from existing information sources in PRACE and implement as many checks as possible in an automatic form. However still several quality tests will require manual operation (e.g. GUI testing or documentation review).

Major components of the certification framework include:

- Service Certification Manager – a component which will be responsible for orchestrating the service certification process by executing the series of tests depending on the particular service by collecting information from existing information sources (e.g. querying INCA), running automatic tests (e.g. checking whether particular service is online, or running more custom shell scripts) and requesting (e.g. through the TTS) performing manual tests.

- Certification Status Presentation – this component will provide a simple web page (e.g. CGI based), which will provide all PRACE users and operational staff with single point of information about the current status of all PRACE services.
- Certification Status Change Notification – this component will allow users and services to subscribe for notification about changes in the certification status (e.g. informing users via email about revoked certificate for a given service).
- Certification Status Database – a central database storing all information related to past and on-going service certifications, current service certification status, quality checks definitions (including custom test scripts), etc.

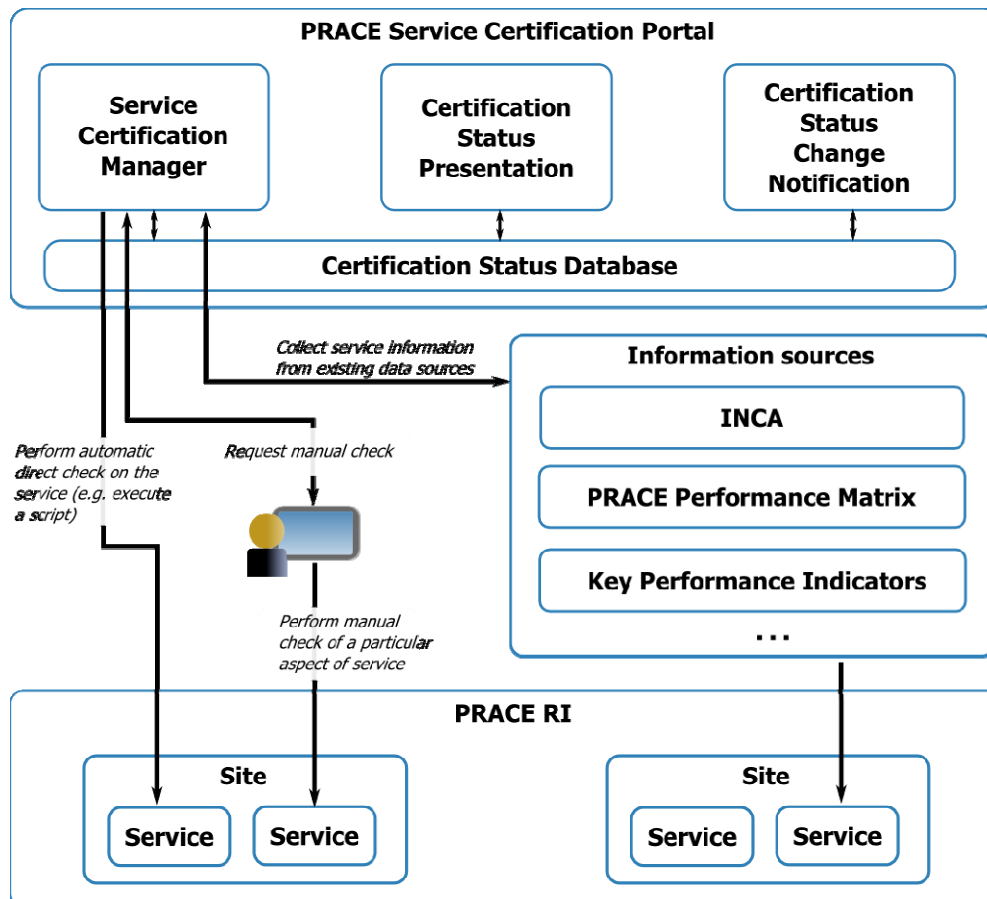


Figure 19: Architecture for PRACE Service Certification platform

Depending on the finalized scope of service certification and selected technologies, several of the components mentioned here could be skipped or merged together, however the bottom-line is to provide PRACE users with a single web page (or document) presenting in a structured form, such as a list or table, information about service status.

Next steps

The service certification activity will continue within the course of the PRACE-2IP project. Procedures will be presented to the Operational Team and project coordination bodies for approval. A tentative work-plan is presented in the table below:

Activity	Milestone
Initial version of the quality checklists and service certification procedure for selected core services in the Wiki	31 May 2012 (Completed)
Discussion of the service certification quality checklists and procedures with service area leaders and collecting feedback. Approval of defined procedures.	30 June 2012
First implementation of the quality checklists and portal system, testing of quality checklists on production services	30 August 2012
Integration of the service certification framework with PRACE services (e.g. INCA), improvement of the quality tests	31 September 2012
Production version of the service certification framework deployment	30 October 2012

Table 8: Service Certification implementation work-plan

5.3 PRACE Information System

Currently PRACE does not provide any real-time information describing the status of computing systems neither to public domain nor to users. The DEISA project was providing users with service and resource availability information using customized views based on Inca monitoring application. However this proved to be inefficient for DEISA staff maintaining the customized Inca views as well as for users as shown by collected user feedback [41]. DEISA users often were not able to extract information necessary to support their work as too much data with an excessive level of details was provided. Furthermore, it was not possible to integrate tools used for collection and management of availability, monitoring and accounting information due to nonstandard technologies and interfaces used.

Considering the importance of service availability and functionality information to the users and user requests for streamlined mechanism to access this data, during the final stage of the PRACE-1IP, Task 6.3 team addressed these requirements and outlined a proposal towards a central PRACE Information System that could be taken over in PRACE-2IP.

PRACE partners implement their own information tools addressing various topics such as system usage [30][31], maintenance messages [32], software availability [33][34]. Since PRACE users are also users of a specific site, they can benefit from services particular to the site they are using. This results in a gap for all users at the time they want to monitor their usage and/or get information status about PRACE Infrastructure. This gap can be filled by a central information system, which can create a simple but common central monitoring service. In this way users can have at least a minimal set of services to check the status of the infrastructure and the usage they are doing.

Within PRACE, information about status of services is internally produced and consumed for operational activities, except for resource usage that provides information about the allocation for a specific project/user.

There are four kinds of information providers currently running, each one covering a specific information domain. Table 9 lists them adding a small description about how they work and which kind of technology is used for exchanging data.

Information	Delivery	Domain	Behaviour	Data exchange
Software/Service availability	WEB [35]	Private (Staff Only, X509 certificate required)	Software Inca: automatic execution of tests (reporters) periodically at each site, Central repository.	XML
Network availability and performance	WEB [36]	Private (Staff Only, X509 certificate required)	iPerf and common tools (ping, fping, traceroute) mainly used. Commands executed at each site, output centrally collected	Unstructured data, socket stream
Resource Usage	DART Client / CGI-based webs hosted at each site	Trusted users (Staff, Users, X509 certificate required)	Each site collects usage statistics of computing resources periodically and on a user/project basis. Info locally stored in a database. Publishing available via CGI and/or a software client (DART). Authorization mechanisms in place and based on roles (user can access only its own data, site admin all data of a site, etc...).	Database, XML
Maintenance	PRACE WIKI	Private (Staff Only, X509 certificate required)	A “maintenance table” is available in the PRACE Wiki. Each site is responsible to add a maintenance entry with the description and duration of maintenance.	HTML, REST

Table 9: List of the PRACE information providers

As shown in Table 9, a lot of work has been done to set up a complete monitoring infrastructure even though the four listed components work in an autonomous way and employ different strategies and technologies for exchanging data. This is a critical point towards a unified platform able to collect all data. A special attention should be also paid to the “Resource Usage” component because the current implementation may change in the near future leaving the floor to the Grid-SAFE framework [38].

Instead of working on interfaces and software workarounds to have all data together from the actual information providers, an alternative and suitable strategy would be to follow a top-down approach, i.e. start defining a small set of information and allowing future enhancement iteratively.

Table 10 is just an example of how a basic information mapping could look like.

Domain	Unit	Value	Audience	Source	Comment
Network	Link Availability	Boolean (Yes/No)	Public	System	--
Network	Link Speed	Numeric (Mbps)	Users	System	--
Compute	PRACE Jobs Running	Numeric	Public	System	--

Compute	PRACE Jobs Queued	Numeric	Public	System	--
Operations	Service Availability	Boolean (Yes/No)	Public	System / PRACE	Availability of services (GridFTP, UNICORE, GSI-SSH, etc)
Operations	Maintenance	Text	Public	System	--

Table 10: Basic Information Map

The table above is an example of how to provide basic but relevant information.

In general audience has been shown to be public but all this work needs a final approval by the management board. Information intended for users could be arranged in a strongly suggested PRACE User Portal, where other information could converge (for example an interface to the PRACE HelpDesk).

Sources of this information are usually supercomputers considered to be Tier-0 and/or Tier-1 for PRACE-2IP). “Service Availability” can include also central services that do not depend on single end systems, it is the case for PRACE Door Nodes services and the PRACE Help Desk.

The layout used to publish this information status is suggested to be as simple as possible. A relevant example comes from Google and the way it publishes status of the services portfolio, Figure 20.

**Figure 20: Google Apps Status**

Other examples that fit more to the HPC ecosystem come from the XSEDE User Portal [39] and the NERSC’s resources live status [40].

5.4 Service areas

5.4.1 Network services

The PRACE network services are based on the developments done in the DEISA project. The assessment, selection and pre-deployment of new technologies for the technical evolution of the network services is delivered within PRACE-2IP WP10 Task 10.1.

5.4.2 Data services

The work of this sub-task has been mainly focused on covering the following data related strands: enhance the offering of data transfer tools; lay the foundation to develop a data management strategy through the assessment of technologies, habits, procedures employed within PRACE sites.

Enhance the offering of data transfer tools

Considering many DECI use cases and reviewing user survey outcomes, it emerged that users often face problems trying to transfer data between their workstation and PRACE machines. This happens because neither there is a way to guarantee their transfers complete successfully, nor there is an agreement on system configurations, policies. For instance, some sites enforce CPU-TIME limits on interactive jobs, such as data transfers, interrupting any activity which exceeds those limits. In order to extend the PRACE offering and improve the reliability of data transfer tools, three technologies were evaluated. Apart the two technologies selected during the first year of the project, UFTP and GlobusOnline, a third one was added: gtransfer. In the following, a description of selected technologies, results of performed tests and a comparison table of their characteristics are presented.

UFTP[14]: The UNICORE File Transfer Protocol is a software component developed within the UNICORE Forum. Its aim is to provide an efficient and reliable tool for data transfer integrated in UNICORE, but also available as a standalone server. At the moment the installation process is a bit complicated and the support of developers is fundamental for deploying and configuring the server; however valuable tests were conducted with success anyway.

The software has been evaluated with respect to two main utilization scenarios: a user who wants to transfer data from his workstation to a PRACE site or vice versa (the so called “last mile”) and a user who wants to transfer data across two different sites.

To reproduce the two utilization scenarios, large data sets have been moved from CINECA to JUELICH (client-server) and between JUELICH and BSC (server-server). In the first case, the PRACE 10Gb/s network connection was used while in the second the tests have been performed using the public internet. These were the only resources available at that time to conduct tests.

Test results showed that the performance of the UFTP is comparable with that of the GridFTP (see **Figure 21**). The time which is necessary to move a 1GB file between CINECA and JUELICH using UFTP is compared to the time which is necessary the same file using GridFTP in the same network conditions. The comparison is executed for different number of streams (i.e. parallel transfer channels). UFTP results slower because of the initial overhead due to establishing the connection. This is due to the communication overhead in UFTP which is introduced at the beginning of the transmission, during the acquisition of the UFTP credentials. This overhead is bigger than the one of GridFTP. A second important feature of UFTP is that it is scriptable and offers a good scheduling functionality. Moreover, UFTP has the valuable feature of being easily integrated in the UNICORE workflow engine.

After this test phase of UFTP, in which it resulted reliable and offering good performances, it will be moved to deployment in PRACE-2IP.

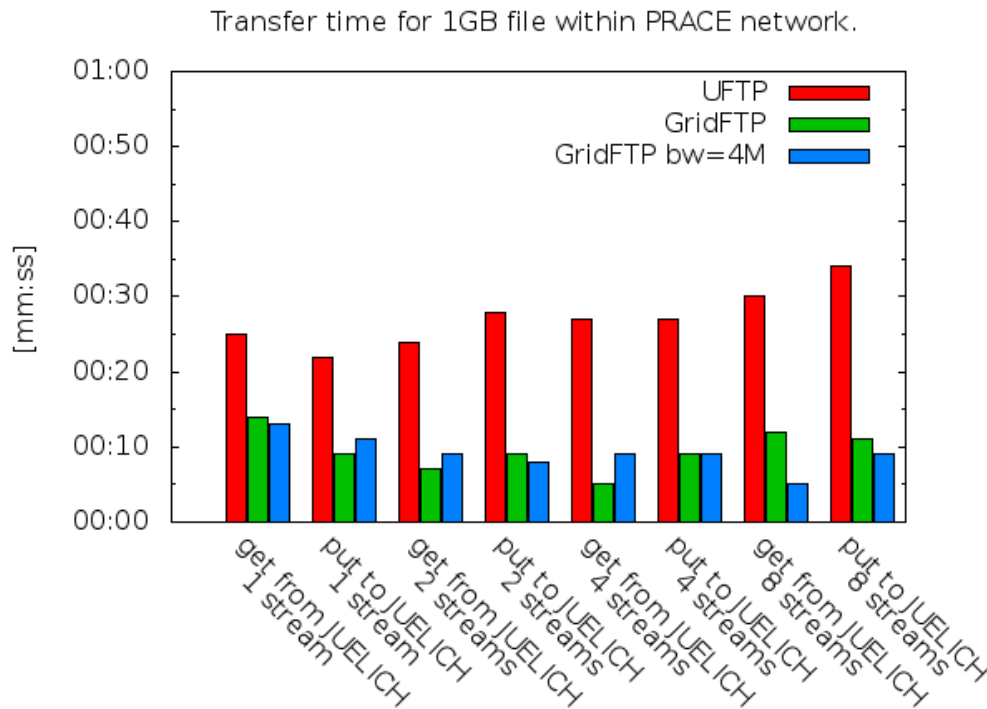


Figure 21: UFTP and GridFTP transfer comparison.

GlobusOnline[18]: The GlobusOnline service is formed of two main components: an on-line service¹ and a GridFTP one-click-installation server, called GlobusConnect. The former automates the tasks associated with moving files between sites, or “endpoints”, while the latter allows a user to transfer files to and from his local machine. The advent of the Globus Connect component has also permitted to install a GridFTP server without requiring administrative privileges or dealing with the details of installing the Globus toolkit.

A preliminary assessment of the service reported some security issues, in particular the need to delegate GSI proxies to an entity being external to PRACE. In fact, in order to “activate” a GlobusOnline endpoint, GlobusOnline needs to act on user’s behalf and thus being able to manage his proxy credentials autonomously. This security issue led the PRACE Security Forum to perform a security assessment to better analyze the tool. The security assessment was supported by one of the GlobusOnline service architect and was considered mandatory for the evaluation of service functionalities to continue. After various discussions, the security constraints have been relaxed enough to permit the evaluation of the service which will take place within PRACE-2IP. However some preliminary tests have been already done and the GlobusOnline service resulted to be very easy and intuitive to use as well as highly reliable. In addition, it brings an interesting feature, called “auto-tuning”, which operates to automatically tune the parameters of the transfer on the base of a continuous analysis of the connection characteristics and transfers historical information. Its actual main limitation is that, being located outside PRACE, it can not handle transfers inside the PRACE private network.

¹ The on-line service is hosted on the Amazon Cloud.

Gtransfer[43] The development of this tool already started in DEISA2 at HLRS and since quite some time it has been publicly available and licensed under the GPLv3. This tool promised to solve two major burdens for the data transfer tasks of our users:

- to remember parameters for optimal high-performance data transfers and to be able to choose more efficient route;
- data movement between different network domains (e.g. Internet vs. PRACE network);

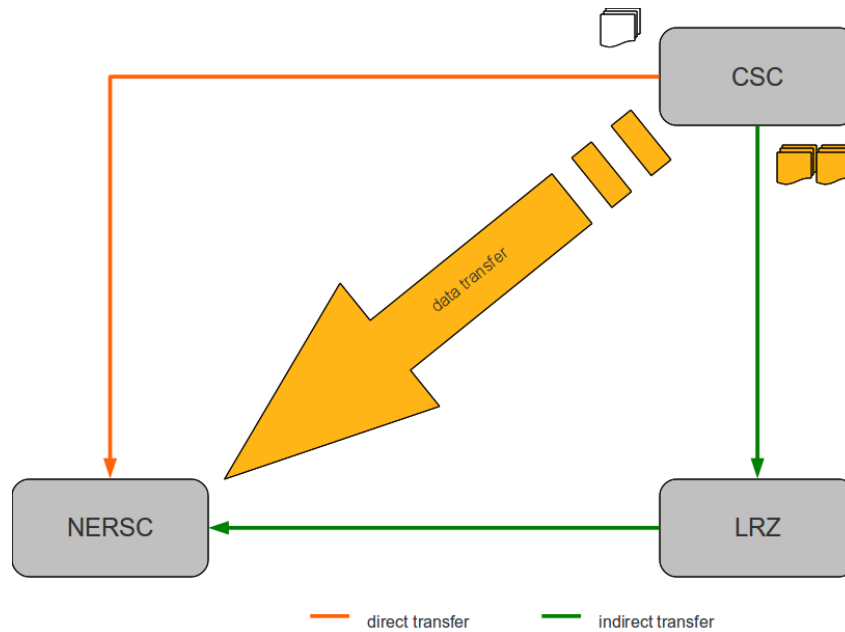


Figure 22: Schema of the *gtransfer* multi path functionality

Gtransfer builds upon the current core data transfer facility in PRACE – GridFTP. It wraps several tools already in use separately to form a new tool providing advanced features to the users. During the evaluation we examined these features of *gtransfer* and evaluated the possible benefit for our users.

Gtransfer can automatically use predefined data transfer parameters, like the number of parallel streams, the TCP buffer size or the number of concurrent transfers for specific connections (solves the first burden). In addition it can perform data transfers crossing network domains by transparently rerouting a data transfer over one or multiple transit/gateway sites (i.e. multi-step transfer which solves the second burden). It also helps the user during command creation by proposing possible options, „known“ hosts (see below) and allows to traverse remote paths directly on the command line.

The evaluation team consisted of members of several PRACE sites (CINECA, CINES, CSC, HLRS and LRZ) providing different test environments including Linux and AIX. CSC was an early adopter of *gtransfer*, as their hosted project Planck-LFI needed to transfer data between NERSC and CSC in an efficient way. Therefore CSC already used *gtransfer* prior to the evaluation.

The performed test activities of the evaluation included:

- installation and configuration of *gtransfer*;
- testing the bash completion feature;
- testing single and multi-step transfers to known (inside the testbed) and unknown (outside the testbed) hosts.

Gtransfer went through this evaluation in a smooth way and no major failures were encountered. Therefore the evaluation report recommends to PRACE to install *gtransfer* as a core tool on each site. The tool will be presented to the PRACE Operation Team in order for being moved to production.

Although, the three evaluation activities described above are currently at three different stages of accuracy, a preliminary comparison of their characteristics was already drafted in order to better focalise next evaluation steps. The comparison is presented in the table below. Nevertheless, it is important to underline that these technologies are not mutually exclusive and could cohabitate as they provide complementary functionalities. In particular, in order to determine which one could be considered the most suitable for addressing a given scenario, they should be weighted with regard to the specific scenario to address.

	GlobusOnline	UFTP	Gtransfer
Performances	Optimal	Almost optimal	Almost optimal
Reliability²	Optimal	Almost optimal	Almost Optimal
Usage easiness	Optimal	Optimal	Almost optimal
GUI available	Yes	Yes	No
Cross-platform	Yes	Yes	No
Control-channel encryption	Yes	Yes	Yes
Data-channel encryption	Optional	Optional	Optional
Authentication	Proxy	x.509 certificate	Proxy
Third-party transfer	Yes	Yes	Yes
Client-server transfer	Yes	Yes	Yes
Multi-path transfer	No	No	Yes
Resources exploitable³	(Selectable) FS	Defined	FS
Open source	No	Yes	Yes
Scriptable	Yes	Yes	Yes

Table 11: Comparison table for the three tools evaluated

² GlobusOnline, exploiting “the cloud”, offers a more modern approach to reliability.

³ With the expression “Resources exploitable” we mean which storage can be accessed by the user. For example, globus-url-copy can access the same file system exploitable by the user, while in GlobusOnline a further parameter can limit it to only a subset of that file system.

Developing a Data Management strategy

Providing the necessary tools for the correct management of data, in particular storing and archiving, was - and still is – one of the challenge of PRACE. On one hand, users accessing Tier-0 systems and generating tens of terabytes of data must be provided with the capability to move, store and archive their data and retrieve them even after the end of the project in an efficient, easy and reliable way. On the other, data management policies and tools should be as similar as possible among PRACE sites to not confuse the users and guarantee a valuable experiencing interacting with the infrastructure.

In order to assess the practices, habits and technologies being adopted at PRACE sites, at least those involved into Task 6.3, a survey was launched and its results collected in a internal white paper. The work based upon previous outcomes or assessment activities which were independently going ahead in other work packages (WP6.3.2b, WP7.6c and WP9.1a). After that, a consolidation and summary of the various results in a single document was produced.

Preliminary, we tried to define a reference model to facilitate the comparison of collected information using a common terminology and semantic for the data management functions of interest for our work. The main idea was to create a comparison table among the most common data management functions, the technologies being adopted by PRACE partners and the user's priorities in order to highlight which uncovered aspects or functions need to be further analysed.

Data management functions considered in our work are listed below.

- **Data movement:** how data is moved across different resources.
- **Data curation:** how data is kept meaningful when archived (for example, how are metadata exploited).
- **Data access:** how user access data (i.e. web-dav, visualization, search, meta-data, etc.).
- **Data preservation:** which policies are implemented for long-term archiving and preservation.
- **Data linking:** how to relate data from different data sets (possibly on different sites).

The resulting comparison is summarized in the following table. The first column presents the data management functions; the second reports the actual PRACE offering; the third shows the technologies which are already in use within PRACE partners but not offered through PRACE; the fourth reports the list of alternative candidate technologies; the last presents the user's priorities (low, medium, high).

Function	PRACE Offering	PRACE Partners	Available Outside	Users priority
Movement	GridFTP (UFTP, gtransfer)	GridFTP, NFSv3, GPFS	UFTP, pNFS/NFV4.1, xrootd	High
Curation	None	iRODS	iRODS, Dspace, Fedora Commons	Low
Access	GridFTP (UFTP, gtransfer)	GridFTP, HDF5, NetCDF, GPFS	iRODS, Dspace, Fedora Commons, UFTP	High

Preservation	None	IBM TSM, HPSS, IBM HSM, dCache	IBM HSM, HPSS, dCache	High
Linking	None	None	PID	Medium

Table 12: Comparison table among data management functions and available technologies

The intention of this work is just to give a picture of the current situation and suggest some recommendations. The data management remains a very complex and heterogeneous topic, but some practical actions could be taken to improve the overall PRACE offering on the data field.

General recommendations that could be derived from the comparison table are listed below.

- The “movement” function is well addressed and the introduction of new technologies, such as *UFTP* and *gtransfer* (evaluated within this task) will soon improve the overall offering.
- The “curation” function is poorly addressed but the corresponding user’s priority is still low.
- The “access” function is well addressed although new technologies for handling specific data format, such as *HDF5*, could be investigated.
- The “preservation” function needs more attention as it results unaddressed while the users’ demand is high. Although there exists a common PRACE agreement for data retention⁴, no common tools to support users are provided. However the preservation of data remains a critical aspect because, on one hand it is difficult for a PRACE site to maintain user data for long period, on the other, transferring large amount of data outside the PRACE Infrastructure might be a challenging task for a user.
- The “linking” function is not addressed by any of PRACE sites while the user’s demand of referring their data sets is growing. Collaboration with other EU projects, such as EUDAT, could help in addressing this point as well as the others.

5.4.3 Compute services

The same methodology focusing on Local Batch Systems and a Uniform Interface for executing jobs over all distributed HPC systems has been followed in the technology watch activity for Compute Services in PRACE-1IP. Advancing in Local Batch Systems has been considered out of scope since they are autonomously chosen and configured by single partners, only recommendations could be made and the only proposal, maybe suitable for the follow up activity in PRACE-2IP, is to create small user communities around specific solutions (e.g. LoadLeveler, PBS, Slurm, etc..) in order to share knowledge about configuration tuning and test of new scheduling algorithms and/or allocation policies. A strong vocation to collaborate is expected to reach valuable results. A good coordination is necessary as well. One among other topics and technology trends that deserve to be covered is related to the cloud computing and how this challenging paradigm can deliver compute services to scientific communities with top level of performance and security and a suitable business model.

⁴ PRACE Contributors Agreement: “Once the allocation of the PRACE User has come to an end, the Contributor must allow the PRACE User access to its data on the Tier-0 System for thirty (30) days. After the thirty (30) day period, the Contributor will save the PRACE User’s data and will hold them for a period of one (1) year up to a reasonable size. After this period the Contributor may erase the data without prior notice.”

Enhancements on UNICORE, which is the only solution adopted for implementing a uniform interface to deliver compute services in PRACE, have been moved to the correspondent operational task (see paragraph 4.4) since they were not related to new features but improvements on the existing ones. Other possible improvements, like the evaluation and a possible implementation in PRACE of a workflow engine has been left as an optional feature, i.e. sites are free to provide it without any support and warranty by PRACE. This because there is not a significant demand gathered from users (see User Survey ran by Task 6.3 during the first year).

Other software solutions, SysFera-DS and ProActive, have been only partially evaluated since they were not considered able to meet the technical requirements of PRACE-1IP. The evaluation of these solutions was driven by their respective vendors which contacted us to present their software.

ProActive Parallel Suite: The ProActive Parallel Suite solution [17] is a java-based middleware for Grid computing. It is developed by INRIA, while an INRIA start-up company, ActiveEon [21], manages and delivers technical support to customers.

ProActive was designed to be a java-based parallel programming model in a grid environment and subsequently extended to resource management, which is what got the PRACE Compute Services area interested. The software component, which is responsible for resources management, is named as “Cloud and Grid IaaS”, formerly known as “Resource Manager”. This component is highly coupled with the other software components and deals with java virtual machines installed at remote sites. In other words, jobs are represented by applications written in java, a Scheduler looks up for the best suitable resource and a Manager is responsible to remotely launch a java virtual machine on the selected host. Users are able to define remote resources where they have access and rights to install and launch JVMs.

The evaluation of ProActive was stopped at very early stage due to an evident mismatch between requirements and features offered. Outcomes are summarized in Table 13. This experience brought up the need to prepare a basic use case in order to help software providers with a clear vision of what the PRACE requirements are.

SysFera-DS: SysFera-DS [22] is a commercial software suite based on the open-source software named “DIET” [23]. Key features of SysFera-DS include an adaptive metascheduling service over heterogeneous resources (Desktop Grid, workstations, commodity clusters, supercomputers, etc...), data management and a workflow engine, software plugins to interact with different Local Batch Systems (LoadLeveler, Torque, Slurm, etc...) and a single sign-on mechanism implemented through a web portal named “SysFera Webboard”.

The underlying infrastructure of SysFera-DS relies on a hierarchical structure based on Master Agents (MA) and Local Agents (LA), which are responsible for gathering information regarding the resources and managing the scheduling actions. Server Daemons (SeD) are responsible, on the other hand, to interact with local resources such as batch schedulers and Cloud platforms as well (see Figure 23).

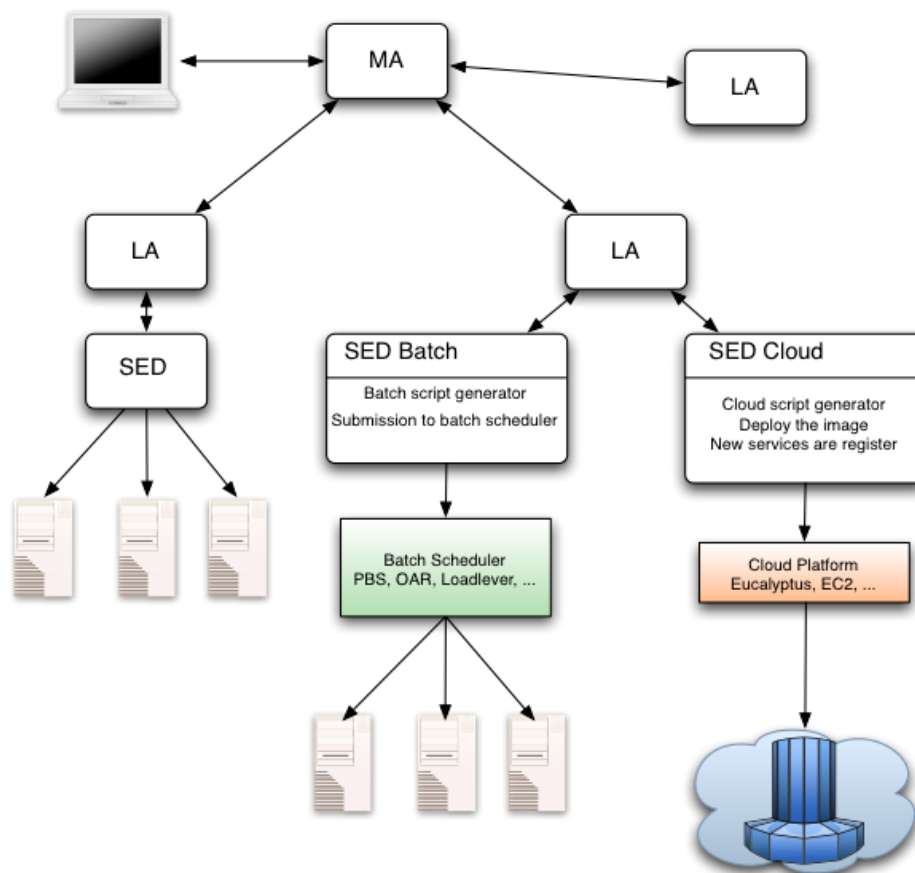


Figure 23: High-level logical design of the software infrastructure implemented by SysFera-DS

Data management algorithms implement the data locality paradigm, like in promising software framework like Google's MapReduce and Apache's Hadoop, to decrease the quantity of data to be exchanged during the staging phase prior the execution of a job.

The test passed a preliminary phase and a demo was used for a further analysis of user functionalities.

One of the strong opportunities offered by SysFera is the big attention to statistics about resource usage and grouped by different domains or dimensions (applications, projects, users, group of users, resources). Figure 24 shows a sample snapshot. This information is critical since it is the basis for developing business intelligence and strategies, which could be very useful both for PRACE project and PRACE AISBL.

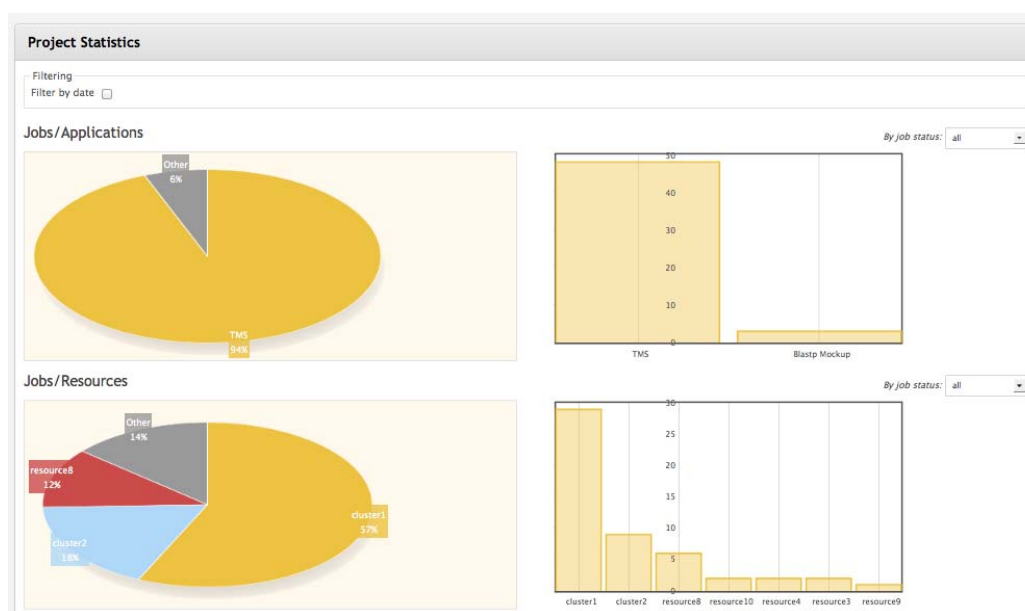


Figure 24: Project statistics page provided by SysFera-Webboard

On the other hand it appeared to be too application-centric even though a workaround is provided for submitting simple job script (which is a basic requirement for PRACE). This is accomplished through a special application named “TMS” (Target Management System).

It was not possible to define the adoption level of this solution by scientific communities and even specific requests have been registered by (potential and actual) PRACE users. The only documented adoption is for the “Décryphon Grid” [24], a joint project with IBM, CNRS and AFM (a French association supporting research in muscular dystrophy) project.

The commercial way for getting support and training is a significant barrier for its adoption.

The following table summarises the evaluation outcomes grouped by strengths and/or opportunities for PRACE and limitations and/or threats detected. Both solutions are not considered to be suitable for delivering compute services within the PRACE Tier-0 infrastructure. The main reason is that both software products address distributed and heterogeneous infrastructures like Grids, where users can get access and execute over a large set of computing systems with an average low level of computing power. This leads to a significant need for having a seamless access to the resources and hiding the underlying heterogeneity. On the other hand, The PRACE Tier-0 infrastructure is established by a small but powerful set of computing resources with an high quality of production services and where users generally prefer to interact directly with a specific batch system in order to gain the maximum benefit of particular architecture.

Software Product	Strengths/Opportunities	Limitations/Threats
ProActive Parallel Suite	<ul style="list-style-type: none"> Supported by different Scientific Communities Abstraction of underlying computing resources 	<ul style="list-style-type: none"> Highly coupled software components Commercial support and training Security (X.509 certificates not supported) Java strongly dependent for user applications Low performance (execution with

		Java Virtual Machines also for native codes) • Suitable for highly heterogeneous Grid environment (Desktop Grid, Cluster, etc..) instead of large supercomputing infrastructures • High maintenance effort is required
SysFera-DS	• Interaction with a large variety of compute services • Statistics and business intelligence • Web interface	• Significant effort required for installation and maintenance • Suitable for highly heterogeneous Grid environment (Desktop Grid, Cluster, etc..) instead of large supercomputing infrastructures • Commercial support and training • Security (X.509 certificates not supported) • Data management policies are not applicable to the actual Tier-0 configuration

Table 13: Evaluation outcomes for ProActive and SysFera software solutions

During the interactions with ProActive and SysFera people and, more generally during the scouting of new technologies in the compute services area, it resulted evident that external providers do not have a clear perception of what the main utilization use cases of the PRACE Tier-0 infrastructure are. To bridge this gap, within the compute services area a very basic and descriptive use case was prepared. The idea is to have a driving document for setting up efficient collaborations focused on real requirements. The following disclaimer statement was to the document due to its informal nature:

“This document has been prepared to support preliminary contacts with external stakeholders (software providers, projects and scientific communities). The views presented in this document don't necessarily reflect those behind the PRACE project and the PRACE AISBL”.

The document, which is available [29], is structured in five sections:

- **PRACE-RI.** Information about what the Research Infrastructure does and a list of the type of calls for users;
- **PRACE-1IP, PRACE-2IP and Work Package 6.** Overview of the two projects currently running and more details about WP6 which is responsible for evaluating and selecting software and/or services to enhance the operational infrastructure;
- **Actors.** Profile of the involved entities who take part of the PRACE business model, e.g. PRACE AISBL, Sites, Systems, Tier-0 and Tier-1 systems and the HPC ecosystem;
- **Business Model.** Linear workflow undertaken from the application form sent to join a particular call to the conclusion of the execution period granted to a project;
- **Conclusions.** Summary of the technical requirements.

The use cases included into the document, were used to guide the preliminary discussion with the vendors of SysFera-DS solution.

5.4.4 AAA services

Enhancement of accounting facilities

Two enhancements of the accounting facilities have been further developed and evaluated by the AAA task:

1. the provision of a central accounting database, based on the Grid-SAFE accounting framework developed by EPCC [13];
2. the provision of budget information to users.

In Figure 25 the addition of the Grid-SAFE based enhancement to the existing accounting facilities is shown in the grey area. Summary usage records are periodically uploaded to the central database. The summaries are for monthly periods. Updates are on a daily basis, so the current month is up-to-date to the previous day.

Figure 25: Accounting architecture with Grid-SAFE facility

A web interface provides access to different reports in different formats: html, pdf, csv and xml. The start page is shown in Figure 26.

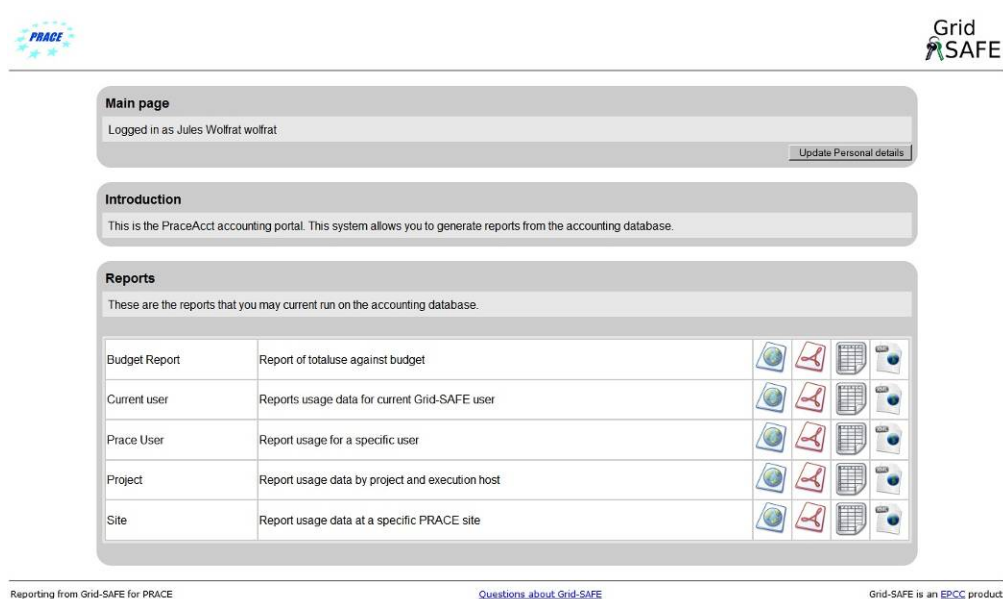


Figure 26: Grid-SAFE client interface

Access is based on the permissions granted to the requestor, using X.509 certificates for the authentication and authorization. Figure 27 shows the report for a particular project in html format.

Usage Report

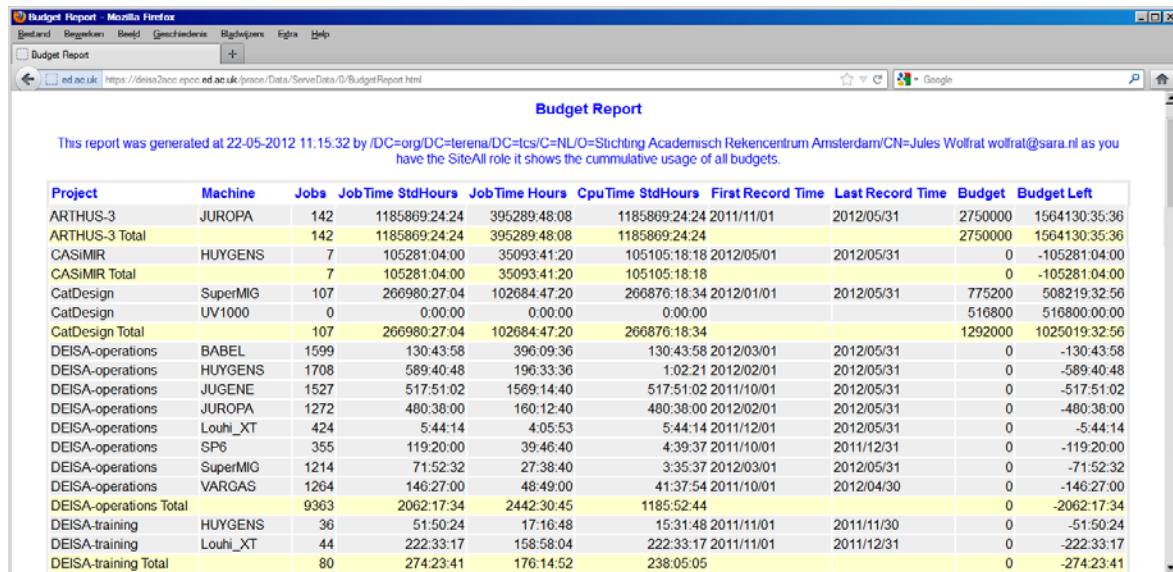
Usage for project MIXTUDI during Jan-2012 - Jun-2012

This report was generated at 22-05-2012 09:49:38 by /DC=org/DC=terena/DC=tcs/C=NL/O=Stichting Academisch Rekencentrum Amsterdam/CN=Jules Wolfrat wolfrat@sara.nl

Project	User	Machine	Jobs	JobTime StdHours	JobTime Hours	CpuTime StdHours	First Record Time	Last Record Time
MIXTUDI	pr1id001	JUROPA	161	675:45:36	225:15:12	675:45:36	2012/03/01	2012/05/31
MIXTUDI	pr1id001	SP6	1207	388513:32:42	129504:30:54	361010:58:06	2012/01/01	2012/05/31
MIXTUDI	pr1id007	SP6	702	97255:32:21	32418:30:47	24364:27:37	2012/01/01	2012/05/31
MIXTUDI Total			2070	486444:50:39	162148:16:53	386051:11:19		

Figure 27: Usage report for a PRACE project

A new feature added this year is that budget information also can be displayed. **Figure 28** shows the budget view for some projects. The right column shows the budget that is available after subtraction of the used resources. DEISA-operations (is PRACE operations now) shows the usage for operational tasks like monitoring. The Grid-SAFE facility soon will be taken in production, after some minor details have been fixed. As can be seen for the operations usage, information from most systems is already available in the repository.



Budget Report

This report was generated at 22-05-2012 11:15:32 by /DC=org/DC=terena/DC=ics/C=NL/O=Stichting Academisch Rekencentrum Amsterdam/CN=Jules Wolfrat wolfrat@sara.nl as you have the SiteAll role it shows the cumulative usage of all budgets.

Project	Machine	Jobs	JobTime	StdHours	JobTime	Hours	CpuTime	StdHours	First Record Time	Last Record Time	Budget	Budget Left
ARTHUS-3	JUROPA	142	1185869:24:24	395289:48:08	1185869:24:24	395289:48:08	1185869:24:24	395289:48:08	2011/11/01	2012/05/31	2750000	1564130:35:36
ARTHUS-3 Total		142	1185869:24:24	395289:48:08	1185869:24:24	395289:48:08	1185869:24:24	395289:48:08			2750000	1564130:35:36
CASIMIR	HUYGENS	7	105281:04:00	35093:41:20	105105:18:18	35093:41:20	105105:18:18	35093:41:20	2012/05/01	2012/05/31	0	-105281:04:00
CASIMIR Total		7	105281:04:00	35093:41:20	105105:18:18	35093:41:20	105105:18:18	35093:41:20			0	-105281:04:00
CatDesign	SuperMIG	107	266980:27:04	102684:47:20	266876:18:34	102684:47:20	266876:18:34	102684:47:20	2012/01/01	2012/05/31	775200	508219:32:56
CatDesign	UV1000	0	0:00:00	0:00:00	0:00:00	0:00:00	0:00:00	0:00:00			516800	516800:00:00
CatDesign Total		107	266980:27:04	102684:47:20	266876:18:34	102684:47:20	266876:18:34	102684:47:20			1292000	1025019:32:56
DEISA-operations	BABEL	1599	130:43:58	396:09:36	130:43:58	396:09:36	130:43:58	396:09:36	2012/03/01	2012/05/31	0	-130:43:58
DEISA-operations	HUYGENS	1708	589:40:48	196:33:36	1:02:21	196:33:36	1:02:21	196:33:36	2012/02/01	2012/05/31	0	-589:40:48
DEISA-operations	JUGENE	1527	517:51:02	1569:14:40	517:51:02	1569:14:40	517:51:02	1569:14:40	2011/10/01	2012/05/31	0	-517:51:02
DEISA-operations	JUROPA	1272	480:38:00	160:12:40	480:38:00	160:12:40	480:38:00	160:12:40	2012/02/01	2012/05/31	0	-480:38:00
DEISA-operations	Louhi_XT	424	5:44:14	4:05:53	5:44:14	4:05:53	5:44:14	4:05:53	2011/12/01	2012/05/31	0	-5:44:14
DEISA-operations	SP6	355	119:20:00	39:46:40	4:39:37	39:46:40	4:39:37	39:46:40	2011/10/01	2011/12/31	0	-119:20:00
DEISA-operations	SuperMIG	1214	71:52:32	27:38:40	3:35:37	27:38:40	3:35:37	27:38:40	2012/03/01	2012/05/31	0	-71:52:32
DEISA-operations	VARGAS	1264	146:27:00	48:49:00	41:37:54	48:49:00	41:37:54	48:49:00	2011/10/01	2012/04/30	0	-146:27:00
DEISA-operations Total		9363	2062:17:34	2442:30:45	1185:52:44	2442:30:45	1185:52:44	2442:30:45			0	-2062:17:34
DEISA-training	HUYGENS	36	51:50:24	17:16:48	15:31:48	17:16:48	15:31:48	17:16:48	2011/11/01	2011/11/30	0	-51:50:24
DEISA-training	Louhi_XT	44	222:33:17	158:58:04	222:33:17	158:58:04	222:33:17	158:58:04	2011/11/01	2011/12/31	0	-222:33:17
DEISA-training Total		80	274:23:41	176:14:52	238:05:05	176:14:52	238:05:05	176:14:52			0	-274:23:41

Figure 28: Budget information for some projects

A new version of DART has been developed too with the feature to display budget information. This is not ready for production yet, but will be further developed in 2IP-WP10. With DART the user can display more detailed usage than with the Grid-SAFE repository because reports for short periods can be displayed too, e.g. for a single day.

Improvement of AA facilities

It was planned to evaluate the Security Token Service (STS), a facility to be developed by EMI, if use cases could be identified. With STS security tokens can be translated from one format to another, which enhances the interoperability between different infrastructures using different middleware. No use cases were identified, so no work has been done on this topic.

Evaluation of the user administration service

The PRACE user administration is based on the LDAP technology. No large changes, which would have been evaluated by the WP6.3 task, have been proposed. Only small enhancements were proposed and these have been implemented by the operations task.

5.4.5 User services

Deployment and extension of the Help-desk Service

The configuration of the PRACE Helpdesk will continue to evolve/extend as new PRACE contributors come online. Any additional services deployed in PRACE will also be included in the Helpdesk.

Looking beyond the boundaries of the PRACE distributed infrastructure, in the context of other major European infrastructure and research initiatives, PRACE is not alone in managing and maintaining a user Helpdesk. Other such Helpdesks include those provided by EGI [16] and the MAPPER project. In order to avoid user confusion and to simplify how a user gains support, the number of entry points to these helpdesks should be kept to a minimum.

As a starting step towards building such a federated research support infrastructure, work will continue as part of PRACE-2IP to evaluate and progress the implementation of an interface between the EGI GGUS (Global Grid User Support) Helpdesk [26] system and the PRACE Helpdesk [25]. A method of transferring user tickets seamlessly between the two systems via an RT-exchange mechanism will be examined.

User Documentation

Whilst online user documentation is now available, a simple means for producing downloadable is also a requirement. Users may prefer to work from local copies of documentation rather than online guides. Work is in progress to implement a PLONE interface to the PRACE SPIP based website. Whilst this interface will be transparent to PRACE users, the publishing environment will enable authors of documentation to both edit documentation and publish it in both web and PDF formats with a single click of a button. The PLONE interface will also remove the current need to manually transfer content from the PRACE SVN repository to the SPIP infrastructure.

PRACE Common Production Environment (PCPE)

The growth of multi-core architecture and the increase in complexity of processors has lead to a large increase in the complexity of optimal task placement on HPC nodes. This is particularly true when users need to under-populate nodes to allow for greater memory per code and/or to decrease off-node network contention. For example, the latest Opteron architecture from AMD (known as 'Bulldozer') is made up of three different levels of components: cores, modules (2 cores which share a FP unit and cache), dies (which consist of a number of module, L3 cache and memory controllers) and processors (which consist of a number of dies linked by HyperTransport). Placing parallel tasks on such a machine is non-trivial.

Unfortunately, the mechanisms for controlling task placement are highly architecture and even vendor specific with solutions ranging from command line arguments to the parallel job launcher to using a file that describes the task placement (or even a combination of both). For users, having to learn both a new (complex) syntax and research the particular processor architecture each time they move to a new facility is a large investment of time and so the majority of scientific applications usually suffer a performance penalty due to poor task placement.

We propose to develop a tool to be included as part of the PCPE that will produce batch submission scripts for HPC resources with a best guess at the optimal task placement based on the number of parallel tasks specified and the processor architecture. The tool will be set up in such a way that once a compute resource is described by a set of parameters in configuration files then users will be able to generate pseudo-optimal batch submission scripts with a simple one-line command. The tool will be designed to be easily extensible to the majority of combinations of HPC resource, batch submission system and processor architecture.

We believe that this tool will make it easier for PRACE users to get the most out of their awarded compute time no matter which PRACE compute resource they are using. The implementation of the tool will be carried out as part of the PRACE-2IP WP10 working plan.

5.4.6 Monitoring services

Section 2.3 introduced KPIs (Key Performance Indicators) that were defined as a mechanism to assess state of PRACE infrastructure. A subset of these KPIs addresses operational qualities of services, such as availability and reliability, deployed in PRACE. The following KPIs were defined for this purpose:

- availability of services,
- reliability of services,
- number of service interruptions,
- average duration of service interruptions,

- fraction of services under monitoring.

Information necessary for measurement of these KPIs is collected by Inca, a tool used for user-level monitoring of availability and functionality of PRACE services. Methods based on Inca were implemented to provide mechanisms for collection and evaluation of necessary information. For instance, to compute KPI describing service reliability information from Inca has to be combined with service maintenance information from PRACE Wiki. All test results for a specific service over a given period of time have to be extracted from Inca. Tests executed while the respective service was under maintenance have to be filtered out. Based on the remaining results the total number of tests and the number of failed tests will be counted to compute service reliability factor.

Similar methods were defined for other KPIs listed above. The defined methods were used to implement a proof of concept application for analysis and reporting on the defined KPIs. Results of this work will be used for design and development of PRACE Information System that was introduced earlier in this section.

5.5 Summary of relevant achievements

This section summarizes the most relevant achievement of Task 6.3 during the second year of the project.

- Fostered the collaboration with different EU projects, such as MAPPER, EMI, IGE. A MoU within EMI was prepared.
- Defined a process for the certification of PRACE services.
- Made a preliminary proposal for the implementation of a centralized PRACE Information System.
- Evaluated the following technologies:
 - perfSONAR
 - UFTP
 - Gtransfer
 - GlobusOnline (partially)
 - SysFera-DS
 - ProActive
- Contributed to the second User Survey (November 2011)
- Published an internal white-paper on data management practices and habits within PRACE [37].
- Enhanced the PRACE Accounting System developing new features for the Grid-SAFE portal interface.
- Extended the PRACE Monitoring System to collect measurements for defined KPIs.
- Refined the PRACE HelpDesk configuration.
- Investigated the possibility to develop a tool to be included as part of the PCPE for the optimal task placement within multi-core architecture.

6 Conclusions

The work WP6 has done in this project, has laid the grounds for a sustainable pan-European infrastructure of Tier-0 systems with a common set of services that allow the provision of a single interaction layer for users and a coordinated operational management. The infrastructure is also capable to be extended to other Tier-0 and Tier-1 systems in the near future. The latter is currently being done through the PRACE-2IP project.

The PRACE distributed research infrastructure is operated and presented to the users as a single research infrastructure, allowing the users to use PRACE as seamlessly as possible. This is done by Tier-0 hosting partners working closely together and synchronising service provision and service deployment as much as possible.

A large number of PRACE common services are deployed in the area of compute services, network services, AAA services, user services, and data services that provide a service layer that integrates the various hosting partner Tier-0 services, and makes the PRACE infrastructure much more than just a collection of individual Tier-0 hosting partners and Tier-0 services.

During the course of this project the PRACE common services integrated the Tier-0 services of JUGENE from GSC@FZJ, CURIE from GENCI@CEA, HERMIT from GCS@HLRS, and SuperMUC from GCS@LRZ and is currently integrating FERMI at CINECA. A new Tier-0 service at BSC is expected to be integrated before the end of 2012.

In the process towards the provision of sustainable and reliable PRACE common services of defined and professional quality, we have made significant achievements since the start of the project. Through a clear roadmap with distinct steps to achieve sustainable quality of operational services on the long term, we took the following steps:

1. the definition and implementation of the PRACE Operational Structure through the PRACE Operational Coordination Team, at the start of year 1 in a matrix organisation with site representatives and service category leaders;
2. the definition and agreement of the set of PRACE common services: in year 1 we have created a first version of the PRACE Service Catalogue, in this reporting year 2 we have refined the PRACE Service Catalogue and added also Tier-1 services. The PRACE Service Catalogue describes the PRACE common services, as well as their service classes (core, additional, optional);
3. the definition, agreement and implementation of operational procedures and policies for the service delivery: in year 1 we have described and implemented common procedures for incident and change management;
4. the definition and implementation of a model for user support: in year 1 we have setup a central helpdesk that is locally managed;
5. the definition of a service certification process to verify, ensure, control and improve the quality of services to be deployed newly; in year 1 and 2 we have defined a complete process for service certification;
6. the definition of a starting set of operational Key Performance Indicators (KPIs): in year 2 we have proposed a set of operational KPIs that are currently being discussed among operational partners and are being implemented;
7. measurement of KPIs followed by the definition of service levels for each of the services: this activity is to be taken up by the Operations work package of the PRACE 2IP and 3IP project.

All these steps are a prerequisite for the implementation of a sustainable set of PRACE common services with quality assurance and quality control (see also 2.3 on PRACE Operational Key Performance Indicators). We can conclude that we have made major steps in this process, which will be continued in the PRACE-2IP and PRACE-3IP projects.

7 Appendix A: PRACE Service Catalogue

The PRACE distributed research infrastructure is well on its path to provide a complete set of sustainable services to its users. Service provision to users is currently mainly done by the Tier-0 hosting partners, governed by the PRACE AISBL statutes and the Agreement for the Initial Period. Relations between Tier-0 sites and their users are typically managed through specific User Agreements between them. PRACE AISBL gives advice to the hosting sites on the allocation of compute resources based on the pan-European PRACE Peer Review. For the execution of the peer review and other services such as the PRACE website, PRACE also uses services provided by third parties. Other important services such as user support and operation of the distributed infrastructure are provided by the PRACE-1IP project.

Tier-1 partners provide access to users, governed by the DECI commitments, currently within the Implementation Phase projects.

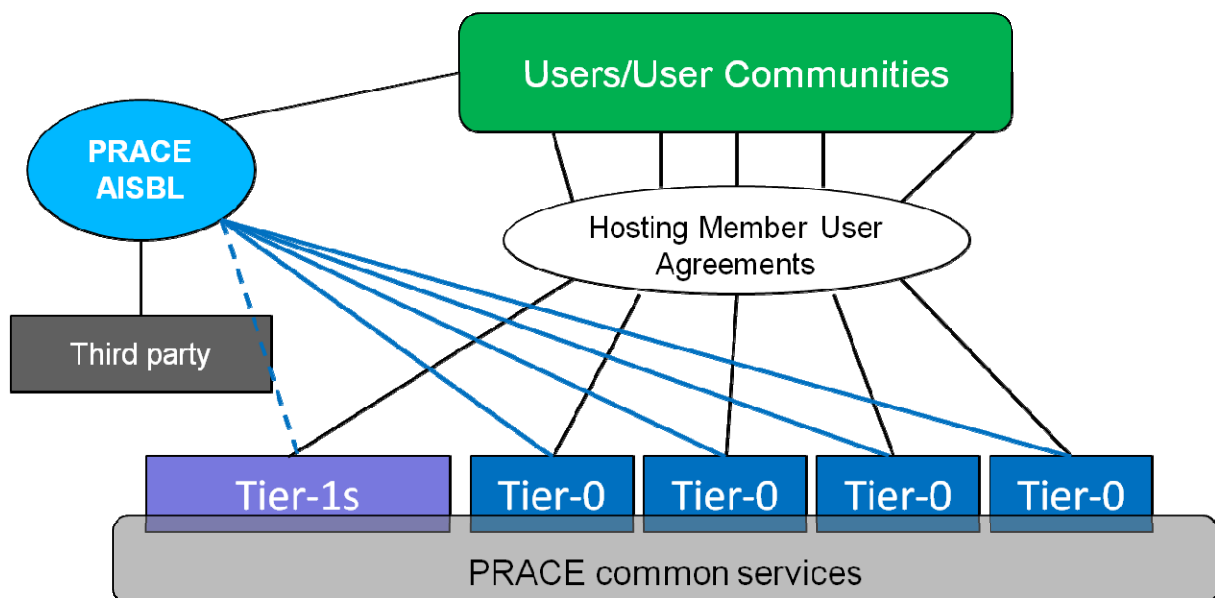


Figure 29: PRACE Service provision scheme and contracts to its users

To support a good and complete overview of all PRACE Operational Services, we have started to develop the PRACE Service Catalogue, which lists and describes the complete set of operational services that the PRACE AISBL is providing, from the point of view of PRACE as a service provider. In addition, Tier-1 services are added to this Service Catalogue to complete the picture of PRACE service provision.

The purpose of the PRACE Service Catalogue is:

- To describe all PRACE operational services
- To define PRACE service categories, and classify all PRACE services accordingly

In this way it describes the full PRACE service portfolio from hosting partners, other partners, the project and the PRACE AISBL.

An important aspect of the PRACE Service Catalogue is the classification of services. We have defined three service classes: Core services, Additional services and Optional services. The availability and support for each of these service classes is defined and described in Table 14.

Core services	
Availability:	Robust, reliable and persistent technologies that must be implemented and accessible at all PRACE Tier-0/1 sites, or provided centrally.
Support:	Support for these services is provided during support hours, i.e. the normal working hours according to the usual working arrangements of the particular Tier-0/1 site.

Additional services	
Availability:	Robust, reliable and persistent technologies that must be implemented and accessible at all PRACE Tier-0/1 sites where possible. Reasons for the service not being implemented at a Tier-0/1 site include technical, legal, financial and policy limitations, whenever an unreasonable effort is needed to provide the service.
Support:	If applicable, support for these services is provided during support hours.

Optional services	
Availability:	Implemented optionally by PRACE Tier-0/1 sites. Availability and long-term support are not guaranteed by PRACE.
Support:	PRACE AISBL and/or Tier-1 partners provide support for these services on a case by case basis, in addition to any support provided directly by the specific site.

Table 14: Classification of PRACE Services as part of the PRACE Service Catalogue

Every PRACE service will be classified according to this classification. It should be noted that the service classes define the availability of the services at the hosting sites, and are not related to service levels.

The definition of the services in the PRACE Service Catalogue is achieved through six criteria:

- **Description:** A brief summary of the service, indicating its value and a general overview of its implementation.
- **Class:** Services are arranged according to their expected availability and support across PRACE hosting partners. This classification is composed of three levels that indicate how essential a service is for the PRACE RI: Core, Additional, and Optional.
- **Provider:** The person(s), group(s), site(s), or team(s) involved in and responsible for the correct implementation and operation of the services.
- **Reference:** Documents and agreements that contain more specific details and information concerning the service provision.
- **Category:** Services are grouped into seven different categories, according to their specific domain: Compute, User, Data, Generic, AAA, Network, and Monitoring.
- **Service:** Concrete services and/or software products that have been chosen to implement the service. For each service/product its Service Class (core, additional, optional) is indicated for Tier-0, Tier-1 and/or PRACE AISBL or a single partner.

The PRACE Service Catalogue will be regularly updated to document the actual status of all services and will be maintained as a living document, where all changes in services and their

provision will be indicated. Status of services can change when new services are deployed, when levels of services are changed, when new service providers (i.e. new hosting partners) are integrated or when new software products are released. The document will at all times reflect the current situation of PRACE services, so that it can be used as the main reference document for service provision within PRACE.

The starting point for the list of services that is listed in the PRACE Service Catalogue has been established in the PRACE-PP in WP4.

This version of the PRACE Service Catalogue v1.6a is the latest and most up to date version at the end of year 2 of this project.

PRACE Service Catalogue V1.6a

Uniform access to HPC				
Description:	Allows a user to execute code on PRACE Tier-0/1 systems, monitor its evolution and retrieve the results across Tier-0/1 systems.			
Class:	Core			
Provider:	Tier-0/1 site + PRACE 1IP/2IP WP6 (compute services representative of the PRACE Operational Team)			
Reference:	Draft User Agreement			
Category:	Compute			
Service:	Product/service	Tier-0	Tier-1	PRACE AISBL or single partner
	UNICORE	core	core	-
	Globus GRAM	optional	optional	-
	Local batch system	core	core	-
	DESHL	-	optional	-
Remarks:	-			

PRACE internal interactive command-line access to HPC				
Description:	Allows a employee of a PRACE partner to connect remotely to a Tier-0/1 system and execute command-line instructions.			
Class:	Core			
Provider:	Tier-0/1 site + PRACE 1IP/2IP WP6 (compute services representative of the PRACE Operational Team)			
Reference:	NA			
Category:	AAA			
Service:	Product/service	Tier-0	Tier-1	PRACE AISBL or single partner

	GSISSH	additional	core	-
	X.509-based SSH	optional	optional	-
Remarks:				

PRACE external (user) interactive command-line access to HPC

Description:	Allows a user to connect remotely to a Tier-0/1 system and execute command-line instructions.			
Class:	Core			
Provider:	Tier-0/1 site + PRACE 1IP/2IP WP6 (compute services representative of the PRACE Operational Team)			
Reference:	Draft User Agreement			
Category:	AAA			
Service:	Product/service	Tier-0	Tier-1	PRACE AISBL or single partner
	At least one of SSH, GSISSH, X.509-based SSH	core	core	-
Remarks:				

Project submission

Description:	Provides Tier-0 users with a centralized point for submitting projects for Peer Review. In case of Tier-1 access, provision of DECI database for project registration.			
Class:	Core			
Provider:	PRACE Peer Review Team			
Reference:	PRACE PP D2.4.2			
Category:	User			
Service:	Product/service	Tier-0	Tier-1	PRACE AISBL or single partner
	PRACE peer review tool (for Tier-0 access)	-	-	core
	DECI database (for Tier-1 access)	-	-	core
Remarks:	-			

Data transfer, storage and sharing				
Description:	Each PRACE User is provided a “home” directory and access to a project space shared with his User Group, at each of the assigned Tier-0/1 sites. The amount of space in each of these directories is indicated in Annex A of the User Agreement for Tier-0 sites. Data can be transferred to and from these directories.			
Class:	Core			
Provider:	Tier-0/1 site + PRACE 1IP/2IP WP6 (data services representative of the PRACE Operational Team)			
Reference:	Draft User Agreement			
Category:	Data			
Service:	Product/service	Tier-0	Tier-1	PRACE AISBL or single partner
	MC-GPFS	optional	optional	-
	GridFTP	core	core	-
	UNICORE	additional	additional	-
Remarks:	GridFTP is a core service for Tier-1 only if a dedicated network is available			

HPC Training				
Description:	Provides training sessions and workshops for topics and technologies in high-performance computing, as well as online and offline education material.			
Class:	Core			
Provider:	PRACE 1IP WP3, PRACE 2IP WP4, Tier-0/1 site, PRACE Advanced Training Centres			
Reference:				
Category:	User			
Service:	Product/service	Tier-0	Tier-1	PRACE AISBL or single partner
	Training portal	-	-	core
Remarks:	-			

Documentation and Knowledge Base	
Description:	User documentation in the form of an online knowledge base, including manuals and other information and tools that are indispensable for the users.
Class:	Core

Provider:	Tier-0/1 site + PRACE AISBL + PRACE 1IP WP6, WP7, WP3 + PRACE 2IP WP6, WP7, WP3			
Reference:				
Category:	User			
Service:	Product/service	Tier-0	Tier-1	PRACE AISBL or single partner
	CMS	-	-	core
	Plone	-	-	core
	DocBook	optional	optional	-
Remarks:				

Data Visualization

Description:	Converts data into images as a tool to help users with analysis.			
Class:	Optional			
Provider:	Specific PRACE sites			
Reference:				
Category:	Generic			
Service:	Product/service	Tier-0	Tier-1	PRACE AISBL or single partner
	Various services and tools	optional	optional	-
Remarks:				

Authentication

Description:	Confirm the identity of a user and bind that user to a new account. This involves identifying a user's certificate, creating a global PRACE RI account for the user on the central LDAP and making it available for distribution on all PRACE RI Resources.			
Class:	Core			
Provider:	Peer Review Team + Tier-0/1 site + PRACE 1IP/2IP WP6 (AAA services representative of the PRACE Operational Team)			
Reference:	PRACE Security Policy			
Category:	AAA			
Service:	Product/service	Tier-0	Tier-1	PRACE AISBL or single partner
	PKI	core	core	-

	MyProxy			core
Remarks:	My proxy is provided by multiple parties (e.g. as backup/disaster recovery).			

Authorization

Description:	Specifies access rights for each user account created based on the content of the specific User Agreement and the PRACE Security Policy. Ensures that security rules and access rights are obeyed, and manages changes to these (based on new security policies or redefined User Agreements).			
Class:	Core			
Provider:	Peer Review Team + Security Forum + Tier-0/1 site + PRACE 1IP/2IP WP6 (AAA services representative of the PRACE Operational Team)			
Reference:	Draft User Agreement, PRACE Security Policy, PRACE Acceptable Use Policy			
Category:	AAA			
Service:	Product/service	Tier-0	Tier-1	PRACE AISBL or single partner
	LDAP (user administration)	core	core	-
Remarks:				

Accounting

Description:	Keeps track of resource usage linked to an account for analysis by users and management. Guarantees that users are not exceeding their limits, as specified by their User Agreement.			
Class:	Core			
Provider:	Peer Review Team + Tier-0/1 site + PRACE 1IP/2IP WP6 (AAA services representative of the PRACE Operational Team)			
Reference:				
Category:	AAA			
Service:	Product/service	Tier-0	Tier-1	PRACE AISBL or single partner
	Apache/CGI DART	core	core	-
	LDAP (user administration)	core	core	-
Remarks:	-			

Information Management				
Description:	Provides a common PRACE collaborative environment for sharing relevant information between PRACE sites (BSCW, wiki, subversion, ...).			
Class:	Core			
Provider:	WP6			
Reference:				
Category:	Generic			
Service:	Product/service	Tier-0	Tier-1	PRACE AISBL or single partner
	TWiki	-	-	core
	SVN	-	-	core
	BSCW	-	-	core
	Prace-ri website	-	-	core
Remarks:	-			

Network Management				
Description:	<p>Establishes and maintains network connections between all PRACE nodes (Tier-0 and Tier-1 systems). The PRACE Network Operations Centre (NOC) operates the PRACE backbone network and the corresponding network monitoring system. The PRACE NOC coordinates networking activities of PRACE partners, who are responsible for creation and management of network connection between the local resources and GÉANT (PRACE backbone).</p> <p>PRACE partner's local network specialists and the PRACE NOC should support PRACE users in using the PRACE network infrastructure.</p> <p>The PRACE backbone will be dedicated, whereas local site connectivity of HPC systems and PRACE servers to the global Internet are public.</p>			
Class:	Core			
Provider:	PRACE NOC and local NOCs of PRACE partners (at least one person per site should be also a network services representative of the PRACE Operational Team)			
Reference:	NA			
Category:	Network			
Service:	Product/service	Tier-0	Tier-1	PRACE AISBL or single partner
	PRACE dedicated network	core	additional	-

	DNS (PRACE RI domain management)	-	-	core
	PerfSonar framework	core	core	-
Remarks:	Dedicated network is an additional service for Tier-1 partners. This means that a dedicated network is required unless unreasonable effort or funding is required. PerfSonar framework is only a service if a dedicated network is available.			

Monitoring

Description:	Periodically presents and analyzes up-to-date essential PRACE parameters and service availability to keep track of the situation of the distributed RI, for example: system uptime/downtime and usage levels, network connections, software and service availability.			
Class:	Core			
Provider:	Tier-0/1 site + PRACE 1IP/2IP WP6 (Monitoring services representative of the PRACE Operational Team)			
Reference:				
Category:	Monitoring			
Service:	Product/service	Tier-0	Tier-1	PRACE AISBL or single partner
	Inca	core	core	-
Remarks:	-			

Reporting

Description:	Periodic reports of system utilization from the Tier-0/1 hosting partner to the PRACE AISBL.			
Class:	Core			
Provider:	PRACE AISBL + Tier-0/1 Hosting Partner			
Reference:				
Category:	Monitoring			
Service:	Product/service	Tier-0	Tier-1	PRACE AISBL or single partner
	Apache/CGI DART	core	core	-
Remarks:	-			

Software Management and Common Production Environment				
Description:	Provides software, tools, libraries, compilers, and uniform mechanisms for software and environment configuration. Presents users with a uniform environment across PRACE Tier-0/1 systems, hiding inessential details such as software installation paths.			
Class:	Core			
Provider:	Tier-0/1 site + PRACE 1IP/2IP WP6 + WP7			
Reference:	NA			
Category:	Generic			
Service:	Product/service	Tier-0	Tier-1	PRACE AISBL or single partner
	Modules	additional	additional	-
	PCPE	core	core	-
Remarks:	-			

First Level User Support				
Description:	Each PRACE User has access to a centrally managed Helpdesk. Issues raised to the Helpdesk are routed to the appropriate First Level Support team. First Level support is responsible for gathering the user's information and determining their issue by identifying what the user is trying to accomplish, analyzing the symptoms and figuring out the underlying problem.			
Class:	Core			
Provider:	Tier-0/1 site + PRACE 1IP/2IP WP6 (User services representative of the PRACE Operational Team)			
Reference:	Draft User Agreement, PRACE 1IP D6.1			
Category:	User			
Service:	Product/service	Tier-0	Tier-1	PRACE AISBL or single partner
	RT-TTS	-	-	core
Remarks:	-			

Advanced User Support				
Description:	Provision of support above and beyond basic problem analysis including but not limited to applications porting, performance tuning, pre-post processing, data access. Higher Level support receives issues that are escalated and routed from First Level User Support.			

Class:	Core			
Provider:	Tier-0/1 site + PRACE 1IP/2IP WP6 + WP7			
Reference:	Draft User Agreement, PRACE 1IP D6.1			
Category:	User			
Service:	Product/service	Tier-0	Tier-1	PRACE AISBL or single partner
	RT-TTS	-	-	core
Remarks:	-			

Service	Service class	Product	Tier-0	Tier-1	PRACE AISBL or single partner
Network management, Monitoring					
Dedicated network	core	PRACE Network	core	additional	
Dedicated network	core	DNS			core
Network management, Monitoring	core	PerfSonar framework	core	core	
Data	-				
Data transfer, storage & sharing	core	MC-GPFS	optional	optional	
Data transfer, storage & sharing	core	GridFTP	core	core	
Data transfer, storage & sharing	core	UNICORE	additional	additional	
Compute	-				
Uniform access to HPC	core	Local batch systems	core	core	
Uniform access to HPC	core	UNICORE	core	core	
Uniform access to HPC	core	Globus GRAM	optional	optional	
AAA	-			-	
Authentication	core	PKI	core	core	
Authentication	core	MyProxy			core
Authorization, Accounting	core	User Administration (LDAP)	core	core	
Accounting, Reporting	core	Apache/CGI DART	core	core	
PRACE internal interactive access	core	GSISsh	additional	core	
PRACE internal interactive access	core	X.509-based SSH	optional	optional	
PRACE external interactive access	core	at least one of SSH, GSISsh, X.509-based SSH	core	core	
User	-			-	
Software management & common production environment	core	Modules	additional	additional	
Software management & common production environment	core	PCPE	core	core	
First level user support, advanced user support	core	RT-TTS			core (tool)
Uniform access to HPC	core	DESHL		optional	
Documentation and knowledge base	core	CMS, Plone, DocBook			core
Project submission, Accounting	core	PRACE peer review tool (for Tier-0 access)			core

Project submission, Accounting	core	DECI database (for Tier-1 access)			core
HPC Training	core	Training portal			core
Data visualization	optional	Various services & tools	optional	optional	
Monitoring	-				
Monitoring	core	Inca	core	core	
Generic					
Information management	core	TWiki			core
	core	SVN			core
	core	BSCW			core
	core	prace-ri website			core

Table 2: Overview of PRACE services, categories and product classes

8 Appendix B: PRACE Operational Key Performance Indicators

14 Operational KPIs, based on ITIL Categories, have been defined:

- (1) Service Availability
- (2) Service Reliability
- (3) Number of Service Interruptions
- (4) Duration of Service Interruptions
- (5) Availability Monitoring
- (6) Number of Major Security Incidents
- (7) Number of Major Changes
- (8) Number of Emergency Changes
- (9) Percentage of Failed Services Validation Tests
- (10) Number of Incidents
- (11) Average Initial Response Time
- (12) Incident Resolution Time
- (13) Resolution within SLA
- (14) Number of Service Reviews

For each of these KPIs, we have defined:

- Description
- Calculation
- Inputs
- Outputs
- Time-interval for measurement
- Tools for measuring the KPI
- ITIL Category for reference
- Implementation plan

Service availability	
Description:	Availability of services
Calculation:	$((A-B) / A) * 100$
Inputs:	Committed hours of availability (A) Outage hours excluding scheduled maintenance (B)
Outputs:	Availability (%)
Time-interval:	Bi-weekly (during every PRACE Operations meeting)
Threshold:	
Tools:	Inca
ITIL Category:	Service Design – Availability Management
Implementation plan:	<p>Inca provides all data necessary for computing this KPI. All test results for a specific service over a given period of time have to be extracted from Inca. Based on the extracted data the total number of tests and the number of failed tests has to be computed. These two numbers should be used in the formula above to compute service availability.</p> <p>The necessary data can be extracted and processed using an SQL query and presented in the PRACE information portal.</p>

Service reliability	
Description:	Availability of services
Calculation:	$((A-B) / A) * 100$
Inputs:	Committed hours of availability (A) Outage hours including scheduled maintenance (B)
Outputs:	Availability (%)
Time-interval:	Bi-weekly (during every PRACE Operations meeting)
Threshold:	
Tools:	Inca
ITIL Category:	Service Design – Availability Management
Implementation plan:	<p>To compute this KPI information from Inca has to be combined with service maintenance information from Wiki or DMOS. All test results for a specific service over a given period of time have to be extracted from Inca, e.g. using an SQL query. Then tests executed while the respective service was under maintenance have to be filtered out. Based on the remaining results the total number of tests and the number of failed tests has to be computed. These two numbers should be used in the formula above to compute service availability.</p> <p>Monitoring data can be extracted and processed using an SQL query. Later this data has to be combined with maintenance information. Depending on its format (Wiki or DMOS) different mechanisms should be used for further processing. When computed, the KPI can be presented on the PRACE information portal.</p>

Number of service interruptions	
Description:	Number of service interruptions.
Calculation:	a) Sum of number of service interruptions excluding scheduled maintenance b) Sum of number of service interruptions including scheduled maintenance
Inputs:	
Outputs:	Interruptions (number)
Time-interval:	Bi-weekly (during every PRACE Operations meeting)
Threshold:	
Tools:	Inca and maintenance information (Inca doesn't indicate always if there is a service interruption; it just can be a problem with the monitoring)

	node). In the maintenance it will (can) be more detailed, although it needs discipline of sites to publish the information.
ITIL Category:	Service Design – Availability Management
Implementation plan:	<p>To compute the first part of the KPI information from Inca is sufficient. For the second part information from Inca has to be combined with service maintenance information from Wiki or DMOS. Based on the test results dataset for a particular service the total number of failure sequences has to be computed. Each sequence is defined by two or more tests, where the first is the first failed test after a passed test and the last is the first passed test after one or more failed tests.</p> <p>It is important to define what do the individual Inca reporters test, as a failure cause (the respective service or monitoring failure) will depend on their functionality. For instance, failed CPE tests might require manual filtering if it can be shown that the respective component works properly.</p> <p>To implement the KPI monitoring data should be extracted and processed using an SQL query. Later, if applicable, this data has to be combined with maintenance information. Depending on its format (Wiki or DMOS) different mechanisms should be used for further processing. When computed, the KPI can be presented on the PRACE information portal.</p>

Duration of service interruptions	
Description:	Average duration of service interruptions.
Calculation:	$SUM(A) / B$
Inputs:	Duration of service interruptions including scheduled maintenance (A) Number of service interruptions including scheduled maintenance (B)
Outputs:	Duration (hours)
Time-interval:	Bi-weekly (during every PRACE Operations meeting)
Threshold:	
Tools:	Inca and maintenance information
ITIL Category:	Service Design – Availability Management
Implementation plan:	To compute the KPI information from Inca has to be combined with service maintenance information from Wiki or DMOS. Based on the test results dataset for a particular service the total number of failure sequences and their duration has to be computed. Each sequence is defined by two or more tests, where the first is the first failed test after a passed test and the last is the first passed test after one or more failed tests. Duration of the respective sequence is defined by the time passed

	<p>between the first and the last tests in the sequence.</p> <p>To implement the KPI monitoring data should be extracted and processed using an SQL query. Later, if applicable, this data has to be combined with maintenance information. Depending on its format (Wiki or DMOS) different mechanisms should be used for further processing. When computed, the KPI can be presented on the PRACE information portal.</p>
--	---

Availability monitoring	
Description:	Percentage of services and infrastructure components under availability monitoring.
Calculation:	$(A / B) * 100$
Inputs:	Number of services under availability monitoring (A) Number of core and additional services in Service Catalogue (B)
Outputs:	Availability monitoring (%)
Threshold:	
Time-interval:	Bi-weekly (during every PRACE Operations meeting)
Tools:	Inca
ITIL Category:	Service Design – Availability Management
Implementation plan:	<p>To compute the KPI a list of services deployed on PRACE resources and a list of monitored services is required. The list of services deployed on PRACE resources should contain all core and additional services available on PRACE resources. For core services it is assumed that such service is either deployed on all PRACE resources or is a central service and is deployed on one or two, i.e. backup, PRACE resources. For additional services an up-to-date list of service deployment is required. The list(s) should be maintained by the respective task/sub-task leaders.</p> <p>The KPI should be computed manually. The current value can be presented on the PRACE information portal.</p>

Number of major security incidents	
Description:	Number of identified security incidents, classified by severity category.
Calculation:	
Inputs:	PRACE Security Forum Incident Reports
Outputs:	Major security incidents (number)

Threshold:	
Time-interval:	
Tools:	PRACE wiki
ITIL Category:	Service Design – Information Security Management
Implementation plan:	<p>The Security Forum has agreed on a set of criteria for a security incident to be labeled as “major”.</p> <p>This will be indicated in a wiki table where all incidents will be logged.</p> <p>See: https://prace-wiki.fz-juelich.de/bin/view/PRACE/Operations/LogofPRACESecurityIncidents </p>

Number of major changes	
Description:	Number of major changes to PRACE services implemented by the PRACE Operational Coordination Team.
Calculation:	See below (implementation)
Inputs:	Major changes included in the “List of Closed Changes” of the PRACE Change Management Tool
Outputs:	Major changes implemented (number)
Time-interval:	Quarterly (once every 3 months)
Threshold:	
Tools:	PRACE Change Management (available in the PRACE wiki)
ITIL Category:	Service Transition – Change Management
Implementation plan:	<p>The Change Management Tool available in the PRACE wiki [1] is used.</p> <p>This KPI is calculated by the sum of all entries included in the section “List of Closed Changes” which satisfy the following conditions:</p> <ul style="list-style-type: none"> • The date value which is part of the multi-value attribute “Status (date of completion)” must be in the considered time interval; • The status value which is part of the multi-value attribute “Status (date of completion)” must be equal to “Implemented” OR “Partially Implemented” • The value of attribute “Type” must be equal to “T62C” OR “T63C” (these values will be soon changed to a more general ones, e.g. “Operations” and “Technology” respectively) <p>[1]: https://prace-wiki.fz-juelich.de/bin/view/PRACE/Operations/ChangeManagement</p>

Number of emergency changes	
Description:	Number of emergency changes to PRACE services.
Calculation:	See below (Implementation)
Inputs:	Emergency changes included in the “List of Closed Changes” of the PRACE Change Management Tool
Outputs:	Emergency changes (number)
Time-interval:	Quarterly (once every 3 months)
Threshold:	
Tools:	PRACE wiki
ITIL Category:	Service Transition – Change Management
Implementation plan:	<p>The Change Management Tool available in the PRACE wiki [1] is used.</p> <p>This KPI is calculated by the sum of all the entries included in the section “List of Closed Changes” which satisfy the following conditions:</p> <ul style="list-style-type: none"> • The date value which is part of the multi-value attribute “Status (date of completion)” must be in the considered time interval; • The status value which is part of the multi-value attribute “Status (date of completion)” must be equal to “Implemented” OR “Partially Implemented” • The value of attribute “Type” must be equal to “URGENT” <p>[1]: https://prace-wiki.fz-juelich.de/bin/view/PRACE/Operations/ChangeManagement</p>

Percentage of failed release component acceptance tests	
Description:	Percentage of release components which fail to pass acceptance tests
Calculation:	$(A / B) * 100$
Inputs:	Number of release components which fail acceptance tests (A) Number of release components tested (B)
Outputs:	Failed acceptance tests (%)
Threshold:	
Time-interval:	Yearly
Tools:	Quality Checklist
ITIL Category:	Service Transition – Service Validation and Testing
Implementation plan:	By the term “acceptance” is meant the certification of a service before entering the production.

	<p>Steps needed for implementing this KPI are:</p> <ul style="list-style-type: none"> • finalize the quality checklists (see Service Certification activity) (https://bscw.zam.kfa-juelich.de/bscw/bscw.cgi/704491); • define an acceptance threshold; • implement the certification process; • measure the KPI.
--	---

Percentage of failed service validation tests	
Description:	Percentage of service validation tests which fail.
Calculation:	$(A / B) * 100$
Inputs:	Number of services which fail validation tests (A) Number of services tested for validation (B)
Outputs:	Failed validation tests (%)
Time-interval:	Quarterly
Threshold:	
Tools:	Quality checklist
ITIL Category:	Service Transition – Service Validation and Testing
Implementation plan:	<p>By the term “validation” is meant the certification of a service which is already production.</p> <p>The implementation plan for this KPI is the same of the previous one (e.g. Acceptance tests).</p>

Number of incidents	
Description:	Number of incidents registered by the Service Desk, grouped into categories
Calculation:	NA
Inputs:	Number of incidents in RT TTS DB
Outputs:	Incidents (number)
Time-interval:	Monthly
Threshold:	
Tools:	RT TTS DB
ITIL Category:	Service Operation – Incident Management
Implementation plan:	Standard TTS query mechanism will be used to extract the data on a monthly basis. Incidents will be categorised as follows:

	Level e.g. Tier1 /Tier0 User Class e.g. Staff /User Area e.g. Network, AAA, Monitoring, User, Data
--	--

Average initial response time

Description:	Average time taken between the time a user reports an Incident and the first time that the Service Desk responds to that Incident.
Calculation:	$SUM(B-A) / C$
Inputs:	Ticket first opened timestamp (B) Ticket creation timestamp (A) Number of incidents (C)
Outputs:	First Response time (hours)
Time-interval:	Monthly
Threshold:	
Tools:	RT TTS DB
ITIL Category:	Service Operation – Incident Management
Implementation plan:	Standard TTS query mechanism will be used to extract the data on a monthly basis. In cases where the response time has not been set, statistics will not be reported. The response time is only set if the ticket is handled in a specific manner by the Helpdesk. The number of missing response times will be monitored and the process checked with a view to minimising the number of occurrences.

Incident resolution time

Description:	Median time for resolving an incident, grouped into categories.
Calculation:	$MEDIAN(A - B)$
Inputs:	Ticket close timestamp (A) Ticket creation timestamp (B)
Outputs:	Resolution time (hours)
Time-interval:	Monthly
Threshold:	
Tools:	RT TTS DB
ITIL Category:	Service Operation – Incident Management
Implementation plan:	Ticket resolution time will be extracted from the TTS on a monthly basis using standard TTS query mechanism.

	<p>The resolution will only take into consideration working days. Working days can be defined as Monday-Friday 0900-1700. National holidays should also be considered but will be treated on a case by case basis as these vary across partners. National holidays will not be defined, but any tickets which fail to meet the KPI will be checked to understand if a national holiday was the reason for the exception.</p> <p>The categories will be defined as follows:</p> <p>Level e.g. Tier1 /Tier0 User Class e.g. Staff /User Area e.g. Network, AAA, Monitoring, User, Data</p>
--	--

Resolution within SLA	
Description:	Rate of incidents resolved during solution times agreed in Contributor's Agreement.
Calculation:	$(A/B)*100$
Inputs:	Number of incidents resolved inside time limit specified in CA (A) Number of incidents (B)
Outputs:	Resolutions within SLA (%)
Time-interval:	Monthly
Threshold:	
Tools:	RT TTS DB
ITIL Category:	Service Operation – Incident Management
Implementation plan:	The rate of incidents resolved within SLA will be calculated on a monthly basis. The resolution time as calculated for the KPI 'Incident Resolution Time' will be used as one of the inputs. This is the resolution time adjusted to take into account the working days/hours in the period.

Number of service reviews	
Description:	Number of formal service reviews carried out during the reporting period.
Calculation:	<ul style="list-style-type: none"> • Number of formal reviews of the PRACE Operations KPIs • Number of formal reviews/updates of the PRACE Service Catalogue
Inputs:	All other KPIs and PRACE Service Catalogue including list of service changes (report on change management)
Outputs:	Service reviews (number)

Time-interval:	4x per year
Threshold:	
Tools:	NA
ITIL Category:	Continual Service Improvement - Service Review
<i>Implementation plan:</i>	Once KPIs are implemented, WP6 (1IP, 2IP, 3IP) will plan 4x per year a dedicated meeting to review all KPIs. One of the meetings will be a f2f meeting, in conjunction with a WP6 (or project wide) all hands meeting.

9 Appendix C: Sample quality checklist for Uniform Access to HPC service

Service Name		Uniform Access to HPC	
Technology		Unicore, Globus GRAM, Local Batch System	
Service Category		Compute	
Class		Core	
Prerequisites			
ID	Check point description	Priority	OK
Accessibility			
	Ensure that Unicore Target sites are available to the authorised users	HIGH	
	IMPLEMENTATION The UNICORE XUADB must be synchronised with the LDAP. Manually procedure: Check that the DN of each user, who has access to the target system, is correctly added in the XUADB. \$XUADB_INST_PATH/bin/admin.sh list dn="\$dn_to_search"		
Compatibility			
	Ensure that list of supported CA (certificate authorities) is compliant with PRACE policy	HIGH	
	IMPLEMENTATION Supported CA trusture is available in the “winnetou” server hosted by SARA: http://winnetou.sara.nl/deisa/certs/keystore.jks		
Documentation			
	A clear information on local batch system type is available to user	LOW	
	A clear information on local file systems and their purpose is present	LOW	
	A clear information on queues available to PRACE user is easily accessible	LOW	
	A clear information on how to use UNICORE in PRACE is available	HIGH	
Security			
	Submission is possible for test user with valid X.509 certificate and appropriate authorization credentials	HIGH	
	Submission is impossible for test user with valid X.509 certificate and lack of appropriate authorization credentials	HIGH	
	Submission is impossible for test user without valid X.509 credential	HIGH	
	Service is accessible through a secure protocol	HIGH	
	Ensure that only the UNICORE Gateway service is world accessible and running as an unprivileged user.	HIGH	
	Ensure that proper CRL lists are configured for Unicore services (check the http://winnetou.sara.nl/deisa/certs/ repository)	HIGH	
	Ensure that only PRACE supported CAs are configured for Unicore	HIGH	
Stress			
	Service must handle at least 1000 simultaneous connections	LOW	
	Service must handle at least 50,000 waiting jobs in the queue	LOW	
	Create 50 jobs and run it asynchronously in batch mode with ucc. IMPLEMENTATION A Python script to create a various number of jobs is available with the UCC package: makebatch.py. First create a directory named “in” and then run: ./makebatch.py 50 samples/date.u In the batch mode only the directory has to be specified and all jobs in this directory will be submitted automatically:		

	bin/ucc batch -i in/ -s <Target Site> -v		
	To remove all jobs from the queue again one could run the following ucc-command: bin/ucc run-groovy -f samples/killall.groovy		
Usability/Functionality			
	Ensure that all required UNICORE services are up and running (Gateway, Unicore/X, XUADB, TSI)	HIGH	
	IMPLEMENTATION <u>Systems using Unicore from EMI distribution:</u> Overall test command: <pre>(service unicore-unicorex status && service unicore-tsi status && service unicore-gateway status) && echo OK echo FAILURE</pre> <p>expected output: last line of the output should contain OK keyword. Otherwise configuration needs to be verified.</p> <u>Other systems:</u> If using unicore from tar distribution, the following recipes are advised: Unicore/X check command: <pre>(ps -ef grep -e java grep -e de.fzj.unicore.uas.UAS)&& echo OK echo FAILURE</pre> Gateway check command: <pre>(ps -ef grep -e java grep -e eu.unicore.gateway.Gateway) && echo OK echo FAILURE</pre> TSI check command: <pre>(ps -ef grep -e perl grep -e tsi) && echo OK echo FAILURE</pre> XUADB check command: <pre>(ps -ef grep -e java grep -e xuadb) && echo OK echo FAILURE</pre> <p>Last line of outputs must contain keyword OK for success.</p>		
	Verify gateway functionality using ucc client	HIGH	
	IMPLEMENTATION: Prerequisites: Determine url of local registry. Example url is as follows: <a href="https://<hostname>:8080/<sitename>/services/Registry?res=default_registry">https://<hostname>:8080/<sitename>/services/Registry?res=default_registry check command: <pre>ucc connect -r <local_registry_url> && echo OK echo FAILURE</pre> <p>expected output: last line of output should contain keyword OK example of valid output: You can access 1 target system(s). OK</p>		
	Check whether connection to local registry is possible	HIGH	
	IMPLEMENTATION Use command: <pre>ucc connect -r <registry></pre>		
	Check for TSI visibility in local registry:	HIGH	
	IMPLEMENTATION		

	Use command: ucc list-sites		
	Check for local storage visibility in local registry	HIGH	
	IMPLEMENTATION Use command: ucc list-storages ucc list-storages -r <local_registry_url> && echo OK echo FAILURE expected output: at least one storage url should be returned by the command AND last line must contain keyword OK indicating successful execution of command		
	Check all above tests with the PRACE Central Registry	HIGH	
	IMPLEMENTATION PRACE Central registry URI: https://prace-unic.fz-juelich.de:9111/PRACE/services/Registry?res=default_registry		
	Verify ability to run jobs via UNICORE native interface by running test job	HIGH	
	IMPLEMENTATION Use command: Ucc run		
	Verify ability to transfer files to UNICORE:	HIGH	
	IMPLEMENTATION Use command: ucc put-file Validate functionality by uploading local file to each of remote storage defined in registry. check command: (repeat for each storage available in local registry) ucc put-file -r <local_registry_url> -s /etc/resolv.conf -t \u6://<storage_name>/PRACE.test && echo OK echo FAILURE expected output: last line of output should contain keyword OK additional check: (verify that file has been successfully transferred) ucc ls -r <local_registry_url> u6://<storage_name>/PRACE.test && OK echo FAILURE		
	Verify ability to run simple job via UNICORE	HIGH	
	IMPLEMENTATION Use command: ucc run samples/date.u -s <Target Site> -v		
	Administration		