SEVENTH FRAMEWORK PROGRAMME
Research Infrastructures

INFRA-2012-2.3.1 – Third Implementation Phase of the European High Performance Computing (HPC) service PRACE

PRACE-3IP

PRACE Third Implementation Phase Project

Grant Agreement Number: RI-312763

D6.1.1
First Annual Operations Report

Final

Version: 1.0
Author(s): Xavier Delaruelle, CEA
Jules Wolfrat, SURFsara
Date: 25.06.2013

Project and Deliverable Information Sheet

PRACE Project	Project Ref. №: RI-312763	
	Project Title: PRACE Third Implementation Phase Project	
	Project Web Site: http://www.prace-project.eu	
	Deliverable ID: < D6.1.1>	
	Deliverable Nature: <DOC_TYPE: Report >	
	Deliverable Level: PU	Contractual Date of Delivery: 30 / 06 / 2013
		Actual Date of Delivery: 30 / 06 / 2013
	EC Project Officer: Leonardo Flores Añover	

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

Document Control Sheet

Document	Title: First Annual Operations Report	
	ID: D6.1.1	
	Version: <1.0 >	Status: *Final*
	Available at: http://www.prace-project.eu	
	Software Tool: Microsoft Word 2007	
	File(s): D6.1.1.docx	
Authorship	Written by:	Xavier Delaruelle, CEA Jules Wolfrat, SURFsara
	Contributors:	Ralph Niederberger (JUELICH), Thomas Bönisch (HLRS), Gabriele Carteni (BSC), Liz Sim (EPCC), Tilo Eissler (LRZ), Gabriele Carteni (BSC), Matteo Lanati (LRZ), Björn Schembera (HLRS), Jutta Docter (JUELICH), Marcello Morgotti (CINECA)
	Reviewed by:	Hank Nussbacher, IUCC Thomas Eickermann, JUELICH
	Approved by:	MB/TB

Document Status Sheet

Version	Date	Status	Comments
0.1	08/04/2013	Draft	Index proposed
0.2	30/04/2013	Draft	First Tier-0 systems input
0.3	07/05/2013	Draft	Section 2, all Tier-0 systems, compute and AAA services
0.4 – 0.5	24/05/2013	Draft	Added network, data, user

			monitoring services, and executive summary
0.6 – 0.7	07/06/2013	Draft for internal review	Content and text formatting harmonization, inclusion of appendix A, removal of helpdesk KPI results.
1.0	20/06/2013	Final version	Integration of internal reviewer comments

Document Keywords

Keywords:	PRACE, HPC, Research Infrastructure

Disclaimer

This deliverable has been prepared by the responsible Work Package of the Project in accordance with the Consortium Agreement and the Grant Agreement n° RI-312763. 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.

Table of Contents

List of Figures

List of Tables

References and Applicable Documents

[1] http://www.prace-ri.eu

[2] PRACE-1IP Deliverable D6.3, Second Annual Report on the Technical Operation and Evolution: http://prace-ri.eu/IMG/pdf/d6.3_1ip.pdf

[3] ITIL KPI information: http://wiki.en.it-processmaps.com/index.php/ITIL_Key_Performance_Indicators

[4] https://www.eugridpma.org/sci/SCI-20130207-v95.pdf

[5] Internet X.509 Public Key Infrastructure Certificate and Certificate Revocation List (CRL) Profile RFC 5280: https://datatracker.ietf.org/doc/rfc5280

[6] EUGridPMA: http://www.eugridpma.org

[7] http://www.globus.org/toolkit/docs/latest-stable/gridftp/

[8] https://github.com/fr4nk5ch31n3r/gtransfer

[9] IGTF: http://www.igtf.net/

[10] DART: http://prace-ri.eu/Accounting-Report-Tool

[11] Globus GSI-OpenSSH: http://www.globus.org/toolkit/security/gsiopenssh

[12] http://prace-ri.eu/Interactive-Access-to-HPC

[13] PRACE User Documentation: http://www.prace-ri.eu/User-Documentation

[14] Inca: http://inca.sdsc.edu/drupal/

List of Acronyms and Abbreviations

AAA	Authorization, Authentication, Accounting
AISBL	Association Internationale Sans but Lucratif (legal form of the PRACE-RI)
AMD	Advanced Micro Devices
AUP	Acceptable Use Policy
BLAS	Basic Linear Algebra Subprograms
BSC	Barcelona Supercomputing Center (Spain)
CA	Certificate Authority
CEA	Commissariat à l'énergie atomique et aux énergies alternatives
CGI	Common Gateway Interface
CINECA	Consorzio Interuniversitario, the largest Italian computing centre (Italy)
CLI	Command Line Interface
CP/CPS	Certificate Policy and Certification Practice Statement
CPU	Central Processing Unit
CSIRT	Computer Security Incident Response Team
DART	Distributed Accounting Report Tool
DDR	Double Data Rate
DECI	Distributed European Computing Initiative
DEISA	Distributed European Infrastructure for Supercomputing Applications. EU project by leading national HPC centres.
DIMM	Dual Inline Memory Module
DP	Double Precision, usually 64-bit floating point numbers
EC	European Community
EGI	European Grid Initiative
EP	Efficient Performance, e.g., Nehalem-EP (Intel)
EPCC	Edinburg Parallel Computing Centre (represented in PRACE by EPSRC, United Kingdom)
EPSRC	The Engineering and Physical Sciences Research Council (United Kingdom)
EUDAT	European Data Infrastructure
EX	Expandable, e.g., Nehalem-EX (Intel)
FDR	Fourteen Data Rate
FFT	Fast Fourier Transform
FZJ	Forschungszentrum Jülich (Germany)
GB	Giga (= 2^{30} ~ 10^9) Bytes (= 8 bits), also GByte
Gb/s	Giga (= 10^9) bits per second, also Gbit/s
GB/s	Giga (= 10^9) Bytes (= 8 bits) per second, also GByte/s
GCS	Gauss Centre for Supercomputing (Germany)
GÉANT	Collaboration between National Research and Education Networks to build a multi-gigabit pan-European network, managed by DANTE. GÉANT2 is the follow-up as of 2004.
GENCI	Grand Equipement National de Calcul Intensif (France)
GFlop/s	Giga (= 10^9) Floating point operations (usually in 64-bit, i.e. DP) per second, also GF/s
GHz	Giga (= 10^9) Hertz, frequency =10^9 periods or clock cycles per second
GigE	Gigabit Ethernet, also GbE
GNU	GNU's not Unix, a free OS
GPFS	IBM General Parallel File System
GPU	Graphic Processing Unit
HDD	Hard Disk Drive

HLRS	High Performance Computing Center Stuttgart (Stuttgart, Germany)
HPC	High Performance Computing; Computing at a high performance level at any given time; often used synonym with Supercomputing
HPL	High Performance LINPACK
IB	InfiniBand
IBM	Formerly known as International Business Machines
ISTP	Integrated System Test Plan
ITIL	Information Technology Infrastructure Library
I/O	Input/Output
JSC	Jülich Supercomputing Centre (FZJ, Germany)
KB	Kilo (= 2^{10} ~ 10^{3}) Bytes (= 8 bits), also KByte
KPI	Key Performance Indicator
LINPACK	Software library for Linear Algebra
LRZ	Leibniz Supercomputing Centre (Garching, Germany)
MB	Mega (= 2^{20} ~ 10^{6}) Bytes (= 8 bits), also MByte
MPI	Message Passing Interface
MPP	Massively Parallel Processing (or Processor)
NAS	Network-Attached Storage
NOC	Network Operations Center
NUMA	Non-Uniform Memory Access or Architecture
OS	Operating System
PB	Peta (= 2^{50} ~ 10^{15}) Bytes (= 8 bits), also PByte
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
PKI	Public Key Infrastructure
PPR	PRACE Peer Review
PRACE	Partnership for Advanced Computing in Europe; Project Acronym
QDR	Quad Data Rate
RAID	Redundant Array of Independent (or inexpensive) Disks
RAM	Random Access Memory
RI	Research Infrastructure
RPM	Revolution per Minute
SAS	Serial Attached SCSI
SATA	Serial Advanced Technology Attachment (bus)
SCI	Security for Collaborating Infrastructures
SHA	Secure Hash Algorithm
SLD	Service Level Description
SSD	Solid State Disk or Drive
SSH	Secure SHell
SURFsara	Dutch national High Performance Computing & e-Science Support Center
TB	Tera (= 2^{40} ~ 10^{12}) Bytes (= 8 bits), also TByte
TFlop/s	Tera (= 10^{12}) Floating-point operations (usually in 64-bit, i.e. DP) per second, also TF/s
Tier-0	Denotes the apex of a conceptual pyramid of HPC systems. In this context the Supercomputing Research Infrastructure would host the Tier-0 systems; national or topical HPC centres would constitute Tier-1
UNICORE	Uniform Interface to Computing Resources. Grid software for seamless access to distributed resources.

Executive Summary

The objective of this report is to present the work that has been done in year one by work package 6, task 6.1, of the PRACE-3IP project on the Operations of the PRACE distributed infrastructure for Tier-0 systems. This work is the continuation of the work done by task 6.1 and 6.2 of the PRACE-1IP project [2].

A new version of the PRACE Service Catalogue, which describes the PRACE common services, has been prepared. The update procedure has been added and two new services have been added as a result of successful evaluations. The new version has been submitted for acceptance by the Tier-0 partners.

In the process towards PRACE Quality of Service and quality control the work on PRACE Operational Key Performance Indicators has been continued. For most of the proposed Indicators the implementation of measurements has been completed and an evaluation will be started in year two.

The Security Forum, responsible for all security related activities, is coordinated by this task. A new version of the Acceptable Use Policy (AUP) for Tier-0 users has been prepared and is submitted to Tier-0 partners for acceptance. Risk reviews of the services Grid-SAFE, gtransfer, gsatellite and UNICORE FTP, have been completed as part of the evaluation of new services.

Based on the procedures for incident and change management the complete set of PRACE common services as defined in the Service Catalogue have been implemented and operated: Network (e.g. the dedicated PRACE network), Data services (e.g., GridFTP), Compute services (e.g., UNICORE), AAA services (e.g., user administration, PRACE Accounting services, GSI-SSH), Monitoring services (e.g., Inca), and User services (e.g., PRACE Common Production Environment, PRACE Help Desk).

Six Tier-0 systems are operational at the end of this project period:

- JUQUEEN at GCS@FZJ;
- CURIE at GENCI@CEA;
- HERMIT at GCS@HLRS;
- SuperMUC at GCS@LRZ;
- FERMI at CINECA;
- MareNostrum at BSC

The systems at CINECA and BSC have been successfully added to the PRACE infrastructure this year and the system JUGENE at GCS@FZJ has been replaced by the JUQUEEN system.

1 Introduction

This report describes the results of Task 6.1 for the first project year of PRACE-3IP. This task is responsible for the operations of the set of common services, which present the PRACE Tier-0 systems as an integrated European infrastructure. It uses the management procedures and organisation as set up by PRACE-1IP. The task further continued the implementation of the roadmap to a professional service level of sustainable services with a defined quality of service.

Section 2 describes the status of the roadmap to sustainable services. This includes the Service Catalogue, which defines the services that are deployed, the operational Key Performance Indicators used and the security activities coordinated by the Security Forum.

Section 3 gives an overview of the Tier-0 systems in use now.

Section 4 gives a status overview of the common services for the different service areas:

- network services (dedicated network provided by GEANT connecting Tier-0 and major Tier-1 centres);
- data services (e.g. GridFTP);
- compute services (e.g. local batch schedulers, UNICORE, Globus GRAM);
- Authorization, Authentication and Accounting (e.g. Public Key Infrastructure (PKI), user administration and accounting, gsi-ssh);
- user services (e.g. first and second level helpdesk, common production environment, user documentation, advanced application production assistance);
- monitoring services for operations.

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 provided as a service layer that integrates the various hosting partner Tier-0 services. This makes the PRACE infrastructure much more than just a collection of individual Tier-0 hosting partners and Tier-0 services.

Service provision to Tier-0 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 managed through specific User Agreements with the sites. PRACE AISBL gives advice to the hosting sites on the allocation of compute resources based on the pan-European PRACE Peer Review (PPR) procedure. For the execution of the peer review and other services, the AISBL 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-3IP project.

The relations among the different stakeholders are shown in Figure 1. Access to Tier-1 services is governed by the DECI commitments of PRACE partners, currently operated by the PRACE-2IP project.

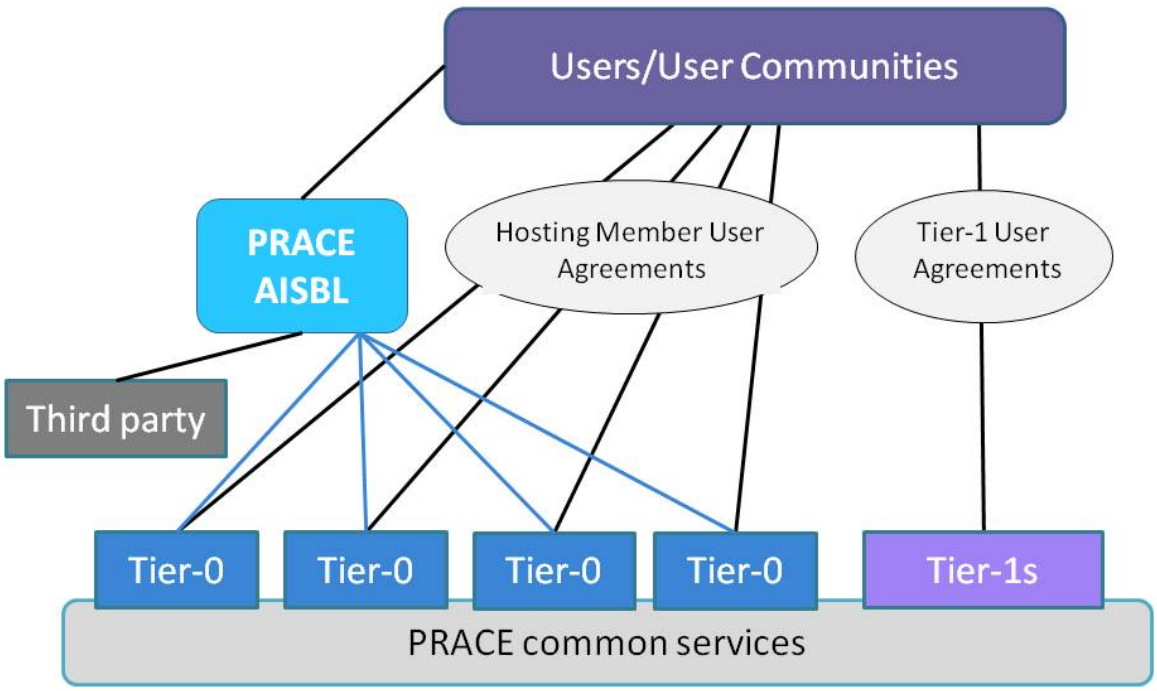

Figure 1: PRACE Service provision scheme and contracts to its users

For the provision of sustainable and reliable PRACE common services of professional quality PRACE-1IP defined a roadmap with distinct steps to achieve a defined quality of service on the long term. This roadmap has been almost completely implemented by PRACE. PRACE-3IP adopted the achievements and continued its further implementation. This roadmap is illustrated by Figure 2 and contains the following steps:

- **the definition and agreement of the set of PRACE common services**: the **PRACE Service Catalogue** describes the services including the service classes used: core, additional and optional. Details are discussed in section 2.2;
- **the definition and implementation of the PRACE Operational Structure**: this has been implemented by PRACE-1IP in a matrix organisation with site representatives and service category leaders. This is the core of the operational coordination team;
- the definition, agreement and implementation of **operational procedures and policies for the service delivery**: this includes common procedures for incident and change management and security policies and procedures;
- the definition of a **service certification** process to verify, ensure, control and improve the quality of services to be deployed newly;
- the definition of a set of **operational Key Performance Indicators** (KPIs): PRACE-1IP proposed a set of operational KPIs and started its implementation. Details are discussed in section 2.3;
- the measurement of KPIs followed by the **definition of service levels** for each of the services: this activity will be started once enough information is available through the KPIs.

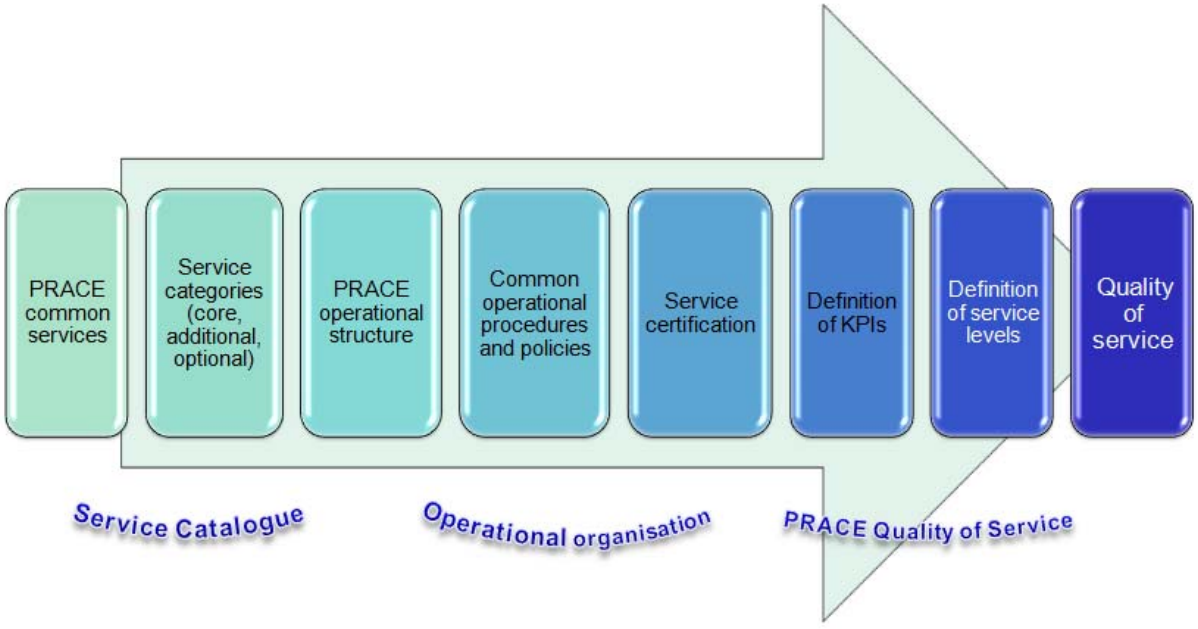

Figure 2: Process towards Quality of Service in PRACE

All these steps are needed for the implementation of a sustainable set of PRACE common services with quality assurance and quality control. To reach this goal, task WP6.1 of PRACE-3IP continued the activities of PRACE-1IP by updating the Service Catalogue and the implementation of the KPIs. The implementation of the service certification process has been the responsibility in this period of WP10 of PRACE-2IP.

Security needs special attention in the PRACE infrastructure, which by its nature must be open for external access by users and services. The Security Forum, set up by PRACE-1IP, continued to coordinate all security related issues and has an important role in delivering reliable services to the PRACE community. Its activities are described in section 2.4.

2.2 Service Catalogue

The Service Catalogue describes the basic set of common services that support the infrastructure. It describes the full PRACE service portfolio provided by hosting partners, other partners, the project and the PRACE AISBL.

An important aspect of the PRACE Service Catalogue is the classification of services. Three service classes have been defined: 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: PRACE Service Catalogue. These 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 is maintained as a living document. 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.

Task 6.1 of PRACE-3IP took over the coordination and maintenance of the Service Catalogue from WP6 of PRACE-1IP. At the end of PRACE-1IP the Catalogue was discussed by the operations team, in which all Tier-0 and Tier-1 partners are represented and was submitted to the PRACE Board of Directors for approval. For several reasons the acceptance was pending and early 2013 it was decided to discuss with the PRACE Hosting Members the open issues. For this discussion a new version has been developed and discussed in WP6 with all partners. The main updates were the addition of two new services, which were accepted within WP6 as a new service in the fall of 2012, and the description of the update procedure of the Service Catalogue. This version is submitted to the Hosting Members for final comments and after acceptance it will be submitted again to the Board of Directors for approval.

2.3 Operational Key Performance Indicators

Quality assurance and quality control are important whenever services are delivered. In the PRACE arena 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 a first and limited set of PRACE Operational Key Performance Indicators (KPIs) has been drafted.

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).

PRACE-1IP defined a limited number of KPIs based on the following criteria:

- Measurable quantitative values;
- Based on ITIL categories [3];
- Give information on the quality of the service & service provision;
- Periodically monitored and registered.

This resulted in the following list:

- Service Availability
- Service Reliability
- Number of Service Interruptions

- Duration of Service Interruptions
- Availability Monitoring
- Number of Major Security Incidents
- Number of Major Changes
- Number of Emergency Changes
- Percentage of Failed Services Validation Tests
- Number of Incidents
- Average Initial Response Time
- Incident Resolution Time
- Resolution within SLD (Service level Description)
- Number of Service Reviews

For each of these KPIs the following categories for description have been defined:

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

An example of such a KPI description is provided in Table 1 below for the Service availability.

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
'KPI Lead':	Guillermo/Gabriele
Implementation plan:	The Change Management Tool available in the PRACE wiki [1] is used. This KPI is calculated by the sum of all entries included in the section "List of Closed Changes" which satisfy the following conditions: - 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

Number of major changes	
	(date of completion" must be equal to "Implemented" OR "Partially Implemented" - The value of attribute "Type" must be equal to "OPERATIONS" or "TECHNOLOGY" ([1]: https://prace-wiki.fz-juelich.de/bin/view/PRACE/Operations/ChangeManagement

Table 1: Example of a complete operational KPI description

The status of this activity is that the measurements for most of the KPIs have been implemented. The implementation of Service Availability, Service Reliability, Number of Service Interruption and Duration of Service Interruptions has been delayed because of the delay in the implementation of the PRACE information portal.

At the start of the second project year, based on the results, realistic service levels, the Service Level Descriptions (SLDs), for the KPIs will be defined. These SLDs will be used to measure the operational performance, the Quality Control, in the second project year.

The results also will be used to review the used set of KPIs and to decide if changes in the list of KPIs are needed.

2.4 Security Forum

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.

All PRACE partners contributing services to the PRACE infrastructure are represented in the forum and can vote on the decisions.

2.4.1 *Security Policies and Procedures*

A new draft for the Acceptable Use Policy (AUP) for Tier-0 users has been produced and sent to the hosting members for acceptance. This document should serve as a common document for all Tier-0 users which will enable the use of PRACE common services without the need of separate documents for each partner. This new draft differs from the current AUP for DECI users, but after acceptance by the Hosting Members this new AUP will also be proposed as a replacement for the document used for DECI users.

Members of the Security Forum participate in the discussions of the Security for Collaborating Infrastructures (SCI) group. This collaboration of several large infrastructures has produced a document which describes a framework for trust between infrastructures [4]. It basically describes a list of policies and procedures which should be implemented. Based on an assessment of available policies and procedures infrastructures can decide to collaborate, e.g. providing mutual access to services. The document is used by the Security Forum to identify missing policies and procedures and hence to further develop the PRACE policies and procedures.

PRACE is a Relying Party member of EUGridPMA [6], the organisation for the accreditation of trusted Certificate Authorities (CAs). Two EUGridPMA meetings (Lyon, September 2012, and Rome, January 2013) have 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.

An important topic currently is the migration from SHA-1 based hash algorithms to SHA-2 algorithms for the signing of the issued end certificates by the CAs. This migration is needed because it is expected that the SHA-1 will be vulnerable in the near future to successful attacks. By the end of 2013 all CAs should issue only SHA-2 signed certificates. For this reason it's important to check that all PRACE X.509 based services will function with SHA-2 signed certificates. The Security Forum coordinates the tests of the involved services, although based on information from the software providers there should be no problems.

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.

Four services have been reviewed in this period:

- Grid-SAFE, a new service for the provision of accounting information. There were no objections to run this service. One partner however has problems exporting the data because of privacy regulations.
- Gtransfer, a new tool for the transfer of data, especially helpful if intermediate hops are needed for the transfer.
- Gsatellite, a new tool for scheduling bulk data transfers.
- UNICORE FTP (UFTP), using the UNICORE infrastructure, especially the security facilities, for the transfer of data.

All these services successfully passed the review. The reviews use a procedure which has been set up by the Security Forum as part of the PRACE-1IP project.

2.4.3 *Operational Security*

All Tier-1 sites are member of the PRACE CSIRT team. Incidents are reported and discussed using an e-mail list and video/telcon facilities. Emergency phone numbers are available for all sites.

In this period two vulnerabilities in PRACE services were reported and handled:

- For GridFTP, a wrong permission could be set if errors were made in the authorisation database.
- For the Apache/CGI script for the retrieval of accounting data at sites a vulnerability was reported and fixed.

There were no security incidents reported by sites in this period.

The PRACE CSIRT team has contacts with the EGI CSIRT for the exchange of information about security incidents which may have an impact on both infrastructures. At the end of April 2013 a face-to-face meeting of the EGI CSIRT team was attended by the Security Forum leader to present the PRACE security activities and to discuss further collaboration. To further strengthen the collaboration it was decided to organize a common security meeting of EGI, EUDAT and PRACE in the fall of 2013.

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. All the six Tier-0 systems that were planned in 2010 are now operational. A detailed technical sheet is provided for each Tier-0 system newly deployed this year.

3.1 Technical overview of current Tier-0 production systems

3.1.1 *JUQUEEN – GCS@FZJ*

The first Tier-0 system JUGENE at the GCS member FZJ has gone out of production in July 2012 and has been replaced by an IBM Blue Gene/Q System named JUQUEEN, managed by the Jülich Supercomputing Centre, Germany.

Figure 3: JUQUEEN

With 393,216 compute cores and a peak performance of 5.0 PFlop/s it was Europe's fastest supercomputer and 5th fastest in the world in November 2012. In early 2013, an extension of 24 racks to 28 racks took place and finally 458,752 compute cores are provided to the user community. With 2 Gflop/s per Watt, JUQUEEN is also one of the most energy efficient systems. As all other supercomputers at JSC, it has access to the fileserver JUST (Jülich Storage Server) providing filesystems via IBM's General Parallel Filesystem (GPFS).

Machine name	JUQUEEN
PRACE partner	GCS@FZJ
Country	Germany
Organisation	JSC
Location	Jülich, Germany
Nature (dedicated system, access to system, hybrid)	Access to system
Vendor/integrator	IBM
Architecture	IBM Blue Gene/Q
CPU (vendor/type/clock speed)	IBM PowerPC® A2, 1.6 GHz, 18 core
CPU cache sizes (L1, L2)	L1: 18x 32KB/32KB I/D, L2: 16x 32 MB on chip
Number of nodes	28,672
Number of cores	458,752
Number of cores per node	16 (compute) +1 (control) +1 (spare)
Memory size per node	16 GB
Interconnect type / topology	IBM proprietary 40 GBps / 5D Torus
Peak performance	5.9 PFlop/s
Linpack performance (measured or expected)	4.8 PFlop/s (expected)
I/O sub system (type and size)	Online Storage: - IBM System Storage DCS3700, DS3512, DS3512 - 5.6 PB + 4.4 PB - RAID6, RAID5, RAID1 Archive Tape System: - 2x STK Streamline SL8500 - 44.5 PB - 16600 Tapes - 48 Tape Drives
File systems (name, type)	Online Storage: - /work, GPFS - / home{x}, GPFS Archive: - /arch{x}, GPFS
Date available for PRACE production runs	November 2012
Link to the site's system documentation	http://www.fz-juelich.de/ias/jsc/EN/Expertise/Supercomputers/JUQUEEN/JUQUEEN_node.html

3.1.2 *CURIE – GENCI@CEA*

CURIE is the second PRACE Tier-0 petascale system, provisioned by GENCI and operated at TGCC (Très Grand Centre de Calcul du CEA) near Paris, France. CURIE has been opened to users on May 2011 and provides now three different kind of computing resources to the PRACE users:

- 5040 Bullx B510 thin nodes with 2 Intel SandyBridge-EP 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-EP (4 cores each) 2.67 Ghz, 2 Nvidia M2090 GPUs, SSD local disk;
- 90 Bullx BCS large nodes with 16 Intel Nehalem-EX (8 cores each) 2.27 GHz, 512 GB of DDR3 memory (4 GB/core), SSD local disk.

All these compute nodes are interconnected with a full fat tree InfiniBand QDR network. The internal Lustre file system of the supercomputer now has a capacity of 3.5 PB and a bandwidth of 150 GB/s and a computing center-wide Lustre file-system has a capacity of 8 PB and a bandwidth of 100 GB/s.

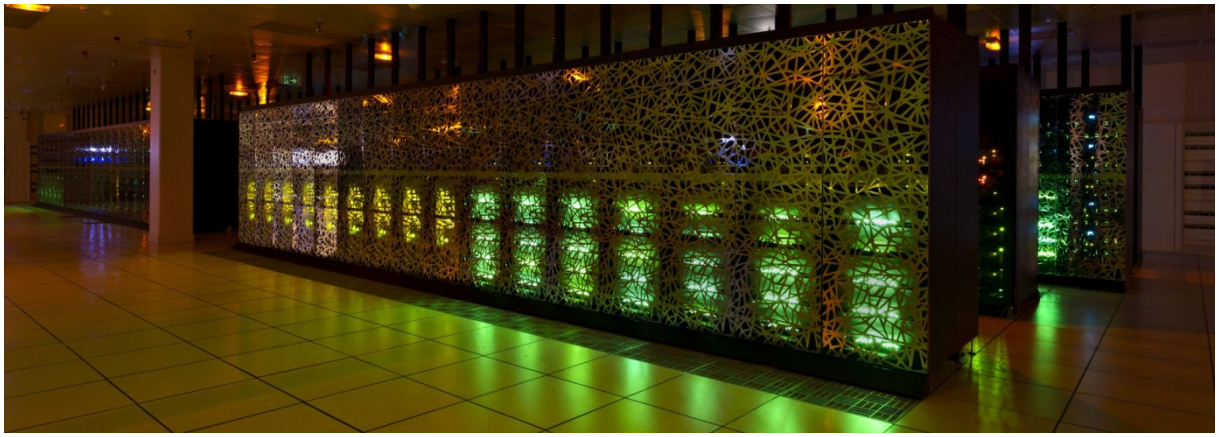

Figure 4: CURIE

3.1.3 *HERMIT – GCS@HLRS*

HERMIT is the 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 (GCS) for PRACE. Delivered in October 2011, HERMIT has been the first PFlop/s system worldwide delivered with AMD Interlagos processors.

Figure 5: 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 grand challenge applications which make use of the whole system or at least large partitions of it in one computational job.

HERMIT provides a maximum performance of 1.045 PFlop/s and a Linpack performance of 831.4 TFlop/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.

Since its deployment in October 2011 HERMIT runs very stable with only few planned maintenances. The operating system is updated continuously.

An extension of HERMIT is scheduled for early 2014.

3.1.4 *SuperMUC – GCS@LRZ*

SuperMUC is the 3 Petaflop/s PRACE Tier-0 system located at the Leibniz Supercomputing Centre (LRZ) in Garching near Munich, Germany, one of the three members of the Gauss Centre for Supercomputing (GCS). GCS brings together the main German national supercomputing centres to provide engineering and computational science with the most advanced European computing infrastructure.

Figure 6: SuperMUC

SuperMUC has been designed by IBM and it delivers its peak performance by means of 155,656 cores grouped in 9,400 nodes. Computing nodes are hierarchically organised in one Fat Node Island and 18 Thin Node Islands.

The Fat Node Island, operational since August 2011 and used as the migration system SuperMIG, is equipped with 10-core Westmere-EX Intel Xeon E7-4870 CPUs. Thin Node Islands are in production since June 2012 and the processing units are 8-core Sandy Bridge-EP Intel Xeon E5-2680 CPUs.

The system is equipped with more than 300 Terabytes of RAM and 4 Petabytes of NAS space for permanent data storage, while the GPFS area for temporary files encompasses 10 Petabytes. The interconnect technology is InfiniBand, with a peak transfer rate of 200 Gigabit per second on the GPFS segment and 80 Gigabit per second on the NAS trunk.

SuperMUC takes energy efficiency very seriously. The local resource manager, IBM LoadLeveler, can handle energy data associated to jobs and adapt the frequency of the processors accordingly. The result is a good trade-off between energy saving and performance loss. The overall environmental impact is reduced through hardware solutions such as "Aquasar", an innovative warm water cooling system for active components developed by IBM. Processors and memories are cooled through pipes filled by water with an inlet temperature of 40° C. This value is significantly higher than the average water temperature of a traditional liquid cooling system, which is around 16° C. This results in a significant energy saving since it is not necessary anymore to spend energy on compressors to cool down the refrigerant. Instead, so-called "free cooling" can be used: the ambient outside air directly cools the hot water from SuperMUC. At the same time, the temperature of the water leaving the system is about 70° C, enough to be used for other processes, such as heating the LRZ building in the winter. All these devices allow keeping the expected energy consumption below 3 Megawatt when SuperMUC provides its peak performance.

3.1.5 *FERMI – CINECA*

FERMI is the new Tier-0 system which is installed and managed by CINECA, Italy. It consists of a 10 rack BlueGene/Q system provided by IBM with a theoretical peak performance of 2.1 PFlop/s. FERMI was delivered for pre-production in August 2012 and went in full production on September 1st, 2012.

Figure 7: FERMI

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. This, together with the 10240 16-cores IBM Power A2 processors, totalling 163840 cores, allows the system to achieve a 1.7 PFlop/s measured performance.

The system is provided with 8 Linux-based frontend nodes. Four of these frontend nodes work as login nodes where users have interactive access and can submit jobs via the IBM Tivoli Workload Scheduler LoadLeveler. The remaining four frontend nodes are used as batch nodes to either perform data transfer and archiving operations or run pre- and post-processing jobs. Two login nodes are directly connected to the 10Gb PRACE Network infrastructure.

I/O nodes are connected to the 3.6PB raw storage system through the high performance InfiniBand network.

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

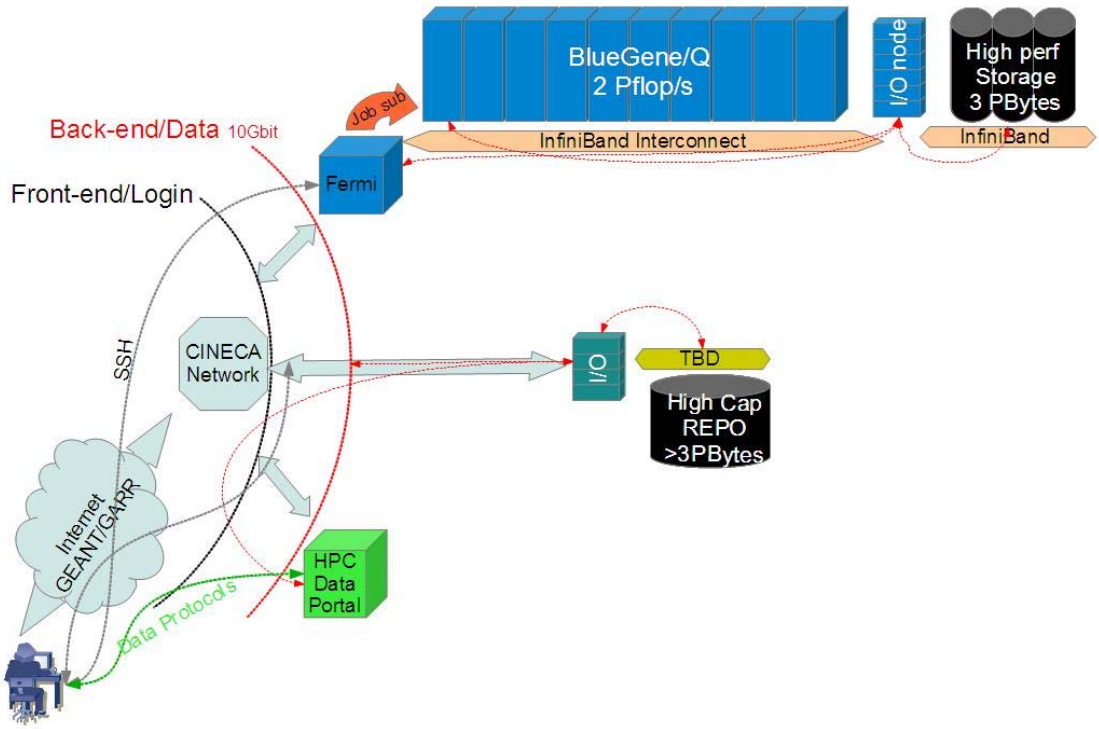

Figure 8: FERMI network and storage schema

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

Machine name	**FERMI**
PRACE partner	CINECA
Country	Italy
Organisation	CINECA
Location	Casalecchio di Reno, Italy
Nature (dedicated system, access to system, hybrid)	Access to system
Vendor/integrator	IBM
Architecture	IBM Blue Gene/Q
CPU (vendor/type/clock speed)	IBM PowerPC® A2, 1.6 GHz, 18 core
CPU cache sizes (L1, L2)	L1: 18x 32KB/32KB I/D, L2: 16x 32 MB on chip

Machine name	FERMI
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	IBM proprietary 40 GBps / 5D Torus
Peak performance	2.1 PFlop/s
Linpack performance (measured or expected)	1.7 PFlop/s
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 - /gpfs/scratch/, SCRATCH filesystem, GPFS - /shared/data, Data repository, GPFS
Date available for PRACE production runs	September 2012
Link to the site's system documentation	http://www.cineca.it/en/hardware/ibm-bgq

3.1.6 *MareNostrum – BSC*

The leading-edge supercomputer MareNostrum hosted and operated by BSC-CNS, Spain, is entered into the third generation phase and it is the latest Tier-0 system integrated into PRACE with a peak performance of one PFlop/s.

It has physically replaced the previous system occupying the same dedicated computer room of about 200 m^2. Despite that it is occupying the same space as its predecessor, the new MareNostrum delivers tenfold more peak computing power, with an increase from 94 TFlop/s to 1000 TFlop/s.

The deployment process is being carried out as planned and is on track. The first part of the machine, with a peak of 700TFlop/s, became available for production at the very beginning of 2013 with the first PRACE user logged on the 2nd of January 2013, coming from a project submitted to the 5th PRACE Regular Call. The integration of the remaining 30% of the system is ongoing and expected to be completed before the summer of 2013.

MareNostrum is based on Intel SandyBridge processors, iDataPlex Compute Racks, Linux Operating System and InfiniBand FDR-10. Key figures can be summarised as follows:

- 36x IBM iDataPlex Compute racks
  - 84x IBM dx360 M4 compute nodes
    - 2x SandyBridge-EP Xeon E5-2670 2.6GHz/1600 20M 8-core 115W
    - 32G DDR3-1600 DIMMs (2GB/core)
    - 500GB 7200 rpm SATA II local HDD
- Total Compute Nodes: 3028
- Total Memory: 94.62 TB
- Peak Performance: 1.0 PFlop/s (Single Node: 332.8 GFlop/s. Single Rack: 27.95 TFlop/s)
- Power Consumption (running HPL Benchmark): 1.08 MW

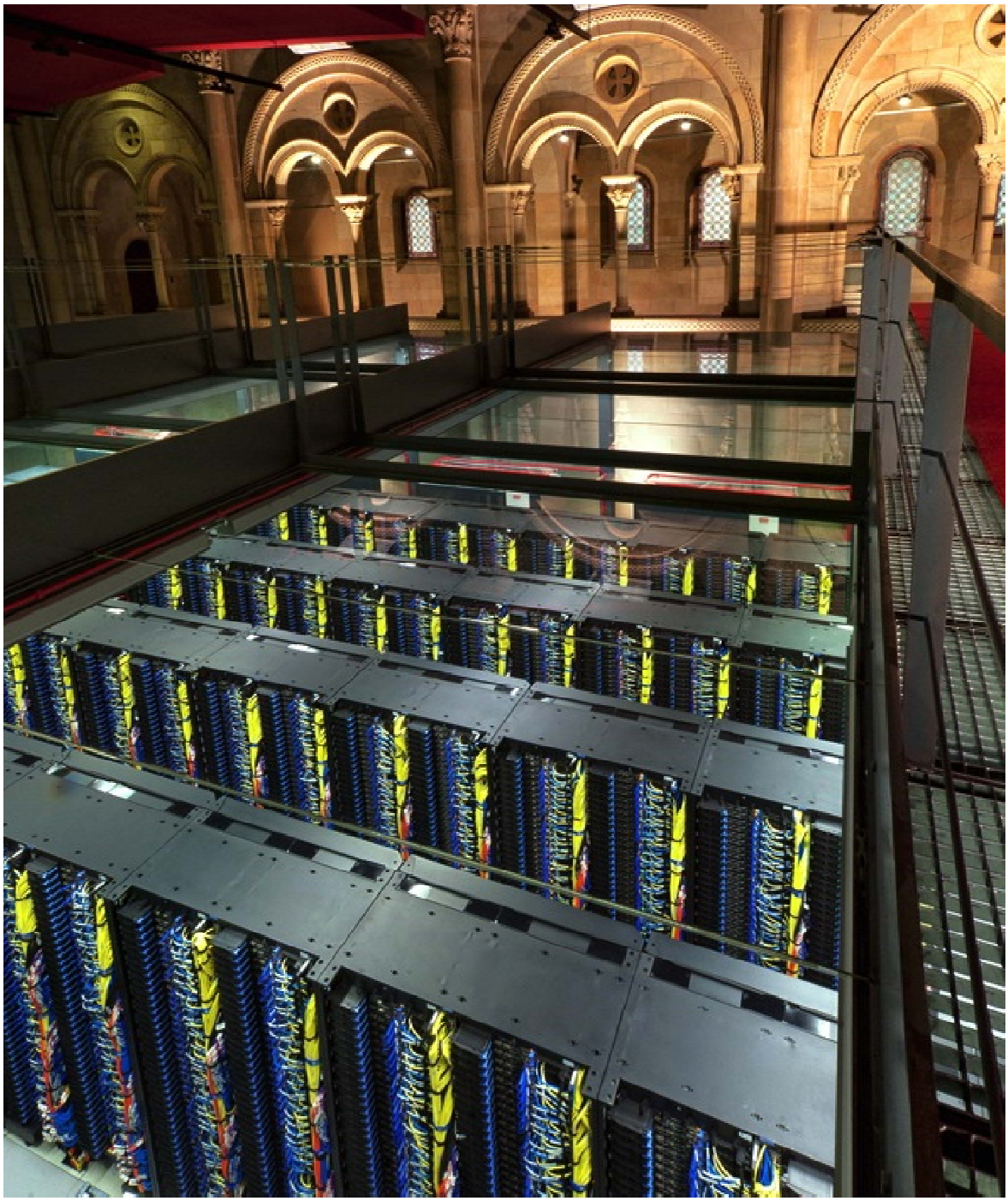

Figure 9: MareNostrum

All compute nodes are connected through an InfiniBand FDR-10 network with a Non-Blocking Fat-Tree topology. The effective bandwidth of 40Gb/s is only intended for high-speed/low-latency communication in parallel applications.

Inside MareNostrum, there are 2 other operating networks, the GPFS Network and the Management Network, both with the same tree-based topology. The GPFS Network, running at 10Gb/s, provides access to the online storage of 1.9PB (2.4PB raw) managed by a GPFS filesystem and secured with a RAID6. The storage mounts all partitions needed for user operations and provides storage services also to the PRACE Tier-1 system MinoTauro. The Management Network, running at 1Gb/s carries the workload of the xCAT Cluster Management Software and all related services (management and distribution of OS images,

Monitoring, Network Boot, etc…). Scheduler and Resource Manager, which rely on IBM LSF, also operate over this network.

An archival/offline storage service for long-term data is also available with 4.1PB of effective capacity (5.45PB raw), in RAID6 and managed with GPFS filesystem as well.

Machine name	**MareNostrum**
PRACE partner	BSC-CNS
Country	Spain
Organisation	BSC
Location	Barcelona, Spain
Nature (dedicated system, access to system, hybrid)	Access to system
Vendor/integrator	IBM
Architecture	IBM iDataPlex DX360M4
CPU (vendor/type/clock speed)	Intel® Sandy Bridge-EP Xeon® Processor E5-2670 2.60GHz 8-core
CPU cache sizes (L1, L2, L3)	L1: 8x 32KB/32KB I/D, L2: 8x 256 KB, L3 (LLC): 20 MB
Number of nodes	3028
Number of cores	48448
Number of cores per node	16
Memory size per node	32 GB
Interconnect type / topology	InfiniBand FDR-10
Peak performance	1 PFlop/s
Linpack performance (measured or expected)	900 PFlop/s (expected)
I/O sub system (type and size)	Online Storage: - IBM System Storage EXP5060 - 1200x SATA 2TB 7.2K rpm - 2.4PB - RAID6 Active Archive: - IBM System Storage DCS 3700/3512 - 180x NL SAS 3TB 3.5” 7.2K rpm - 77x SAS 600GB 3.5” 15K rpm - 5.45PB - RAID6
File systems (name, type)	Online Storage: - /gpfs/home, GPFS - /gpfs/scratch, GPFS - /gpfs/apps, GPFS - /gpfs/projects, GPFS Active Archive: - /gpfs/archive, GPFS
Date available for PRACE production runs	January 2013
Link to the site's system documentation	http://www.bsc.es/marenostrum-support-services/mn3

4 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 published in the PRACE service catalogue and once chosen, the responsibility for a service is taken by the respective service area.

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

4.1 Network services

The PRACE network has constantly evolved over the last years, since its initial deployment by DEISA. As of May 2013, it now connects 20 HPC systems from 18 PRACE partners throughout Europe via a star topology network realized mostly by dedicated 10 Gb/s wavelength through the national NRENs and GÈANT infrastructure.

The main tasks performed have been optimizing, operating and monitoring the network infrastructure, attaching new and detaching old HPC systems to/from the network infrastructure. Since last year, JUGENE Tier-0 system has been decommissioned, JUQUEEN, FERMI and MareNostrum Tier-0 systems have been configured into the network infrastructure and optimized for maximum throughput.

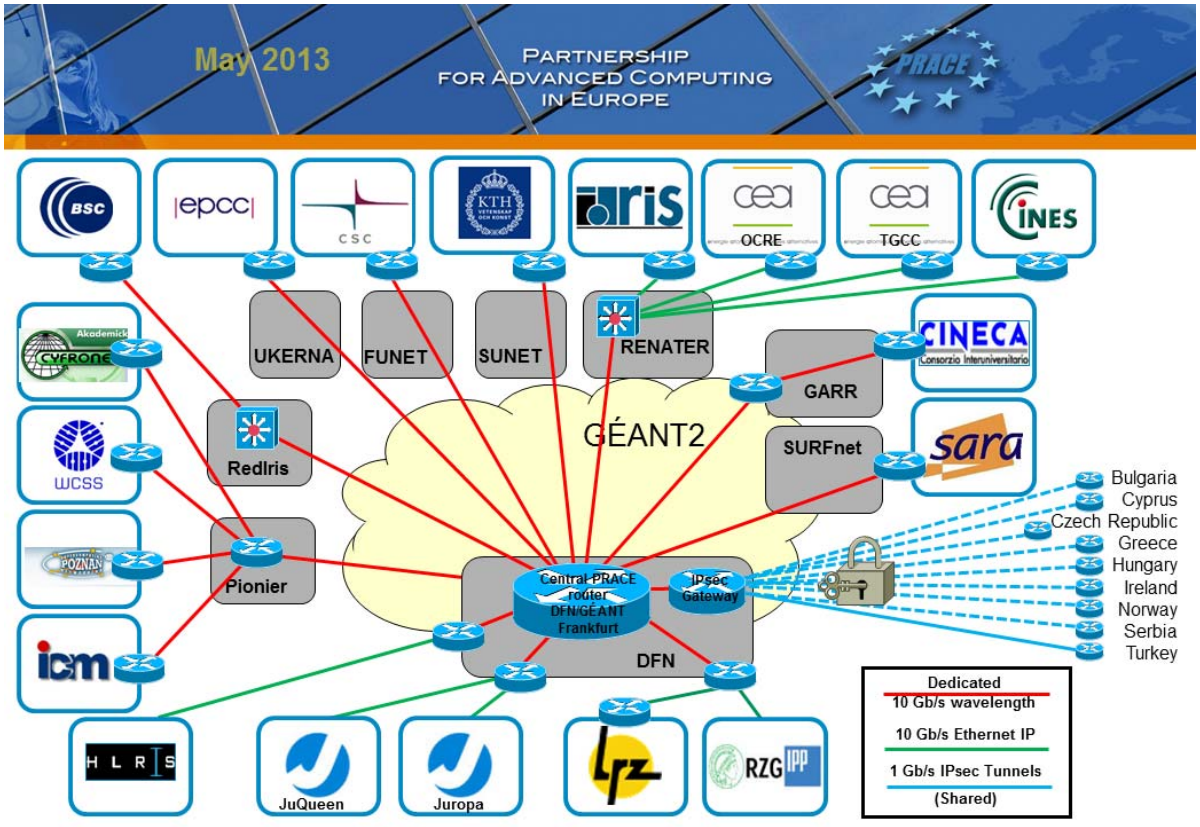

Figure 10: PRACE network layout on May 2013

4.2 Data services

Data services provide methods and tools for data management to fully utilize high-speed connections within the PRACE network and to handle large amounts of data in petascale computing.

To meet these requirements, GridFTP is used as the basic technology for transfers between the centres. Since this technology has been previously used in PRACE-1IP and PRACE-2IP and knowledge has been built around GridFTP, its use continues in PRACE-3IP. GridFTP is a data transfer protocol that supports parallel streams, multi-site transfers, resuming of lost transfers, etc. [7].

A GridFTP server consists of two parts: The frontend is visible for the outside and handles the control connections whereas the backends manage the data transfer itself. We propose a split configuration where the frontend and backend servers reside on distinct hosts. This configuration is safer because the backend servers need to have access to the parallel file systems of the supercomputers and so they are not visible from the outside. Furthermore it is beneficial to have more than two backends (on the same machine) for better performance with parallel data streams.

4.2.1 *Status of Deployment*

In the first year of PRACE-3IP, BSC and CINECA, two sites that host Tier-0 systems for the first time, have deployed the service. In addition, GSC@FZJ has replaced their Tier-0 system and provided a new deployment of GridFTP.

In May 2013 the deployment status of GridFTP is as follows:

Site / Tier-0 System	Version (GridFTP/GT)
CEA / CURIE	6.19/GT5.2.3
CINECA / FERMI	6.14/GT5.2.2
FZJ / JUQUEEN	6.19/GT5.2.4
HLRS / HERMIT	6.19/GT5.2.4
LRZ / SuperMUC	6.19/GT5.2.3

Table 2: GridFTP status at Tier-0 sites

Most sites run the newer versions (5.2.4 and 5.2.3) and no one is running the buggy version 5.2.1. Progress on deployment is maintained on the PRACE internal Wiki.

4.2.2 *Advanced User Tools*

gtransfer is a wrapper tool for the GridFTP CLI that overcomes some disadvantages of the clients and that allows for data transfers with optimized performance and provides an easy and helpful interface to set up data transfers. It is based on GridFTP and uses tgftp (a tool for benchmarking, testing and data transfers which itself uses globus-url-copy (guc)) and uberftp and is developed and supported by Frank Scheiner from HLRS [8].

Since gtransfer has passed the *ISTP* (Internal Specific Targeted Projects) process in late 2012, gtransfer is now listed as an additional service.

In the first year of PRACE-3IP several Tier-0 sites have deployed gtransfer on their systems. In May 2013 the gtransfer deployment status is as follows:

Site / Tier-0 System	Version
CEA / CURIE	v0.0.9b
CINECA / FERMI	0.1.2
HLRS / HERMIT	v0.0.10a
LRZ / SuperMUC	v0.0.10

Table 3: gtranfer depoyment status at Tier-0 sites

It has to be noted that an additional service is not mandatory to be deployed.

4.3 Compute Services

Compute services provide interfaces between users and computing capabilities. Target of this activity is to find out what services can be commonly provided in PRACE, in particular for Tier-0 systems during this first year of PRACE-3IP.

Local Batch Systems are software responsible for managing user jobs. With "job" we not only consider sequential and/or parallel applications but also complex tasks like execution chains or workflow, data processing and data management. This software, which is usually part of a supercomputer software stack, plays a decisive role for finding out a common layer to access different systems and architectures.

The way to interact with batch systems is similar for different products. A scripted file must be edited with a full description of a job, e.g. resources like core/cpu-hours, number of cores/cpu, I/O staging, service level, and all needed action for preparing the execution environment. What differs is the way a job is described, basically made using predefined keywords. Another difference is about features that are provided, which depend on the underlying computing resources, e.g. availability of hardware accelerators, schedulers and application launchers, and shared or distributed memories.

Table 4 gives the list of Batch Systems, which include features for resources management and job scheduling, deployed on all 6 Tier-0 systems currently available. The overview shows the predominance of the IBM Tivoli Workload Scheduler LoadLeveler. Its version for x86–based machines has been withdrawn on March 2013 for marketing reasons and replaced by IBM LSF. Meanwhile it will be still available for IBM BlueGene/Q systems. A common platform is available on two types of systems, iDataPlex and BlueGene/Q.

Site / Tier-0 System	Batch System	Platform
BSC / MareNostrum	LSF	IBM iDataPlex DX360M4
CEA / CURIE	SLURM	Bull Bullx B505/B510/BCS
CINECA / FERMI	LoadLeveler	IBM BlueGene/Q
FZJ / JUQUEEN	LoadLeveler	IBM BlueGene/Q
HLRS / HERMIT	Torque/Moab	CRAY XE6
LRZ / SuperMUC	LoadLeveler	IBM iDataPlex DX360M4

Table 4: Batch Systems on Tier-0 Systems

These similarities make it interesting to develop a set of wrappers with a common syntax for describing a job around the four different batch systems and the four different platforms. Such wrappers will make it easy for users to migrate their work from one system to another.

Work in this direction has been undertaken within task 6.3 of WP6, which has a focus on new technologies and services. Future action for the second year is to evaluate the viability of such an approach.

Another way to create and operate on a common layer is through an even higher abstraction level. This is accomplished by the UNICORE software that is part of the PRACE software portfolio since the beginning and adopted in the DEISA projects. It allows a user to manage single jobs as well as a workflow of jobs remotely through a Java-based graphical interface (URC) and/or a command line client (UCC). Access to end-systems relies on account information stored in the PRACE central LDAP, while resources are published on a central directory service by sites or resource providers.

Table 5: UNICORE software components deployed on Tier-0 Systems shows how different software components of UNICORE are deployed in PRACE.

Component	Description	Deployment
REGISTRY	Directory service publishing Tier-0 resources.	FZJ (Primary), CINECA (Backup)
UNICORE/X	Translate abstract jobs into concrete jobs for a specific target system. Job submission and monitoring	Tier-0 System
GATEWAY	Main entrance to each Tier-0 system. Client connections go through the gateway, which forwards them to internal components, and vice versa.	Tier-0 System
XUUDB	User Database for authentication and authorization. It is synchronised with the PRACE LDAP.	Tier-0 System
TSI	Interface with the local batch system and storage resources	Tier-0 System
URC	Graphical user client (based on Eclipse Java)	Client-side
UCC	Command-line user client	Client-side

Table 5: UNICORE software components deployed on Tier-0 Systems

Figure 11 gives a graphic overview of the deployment.

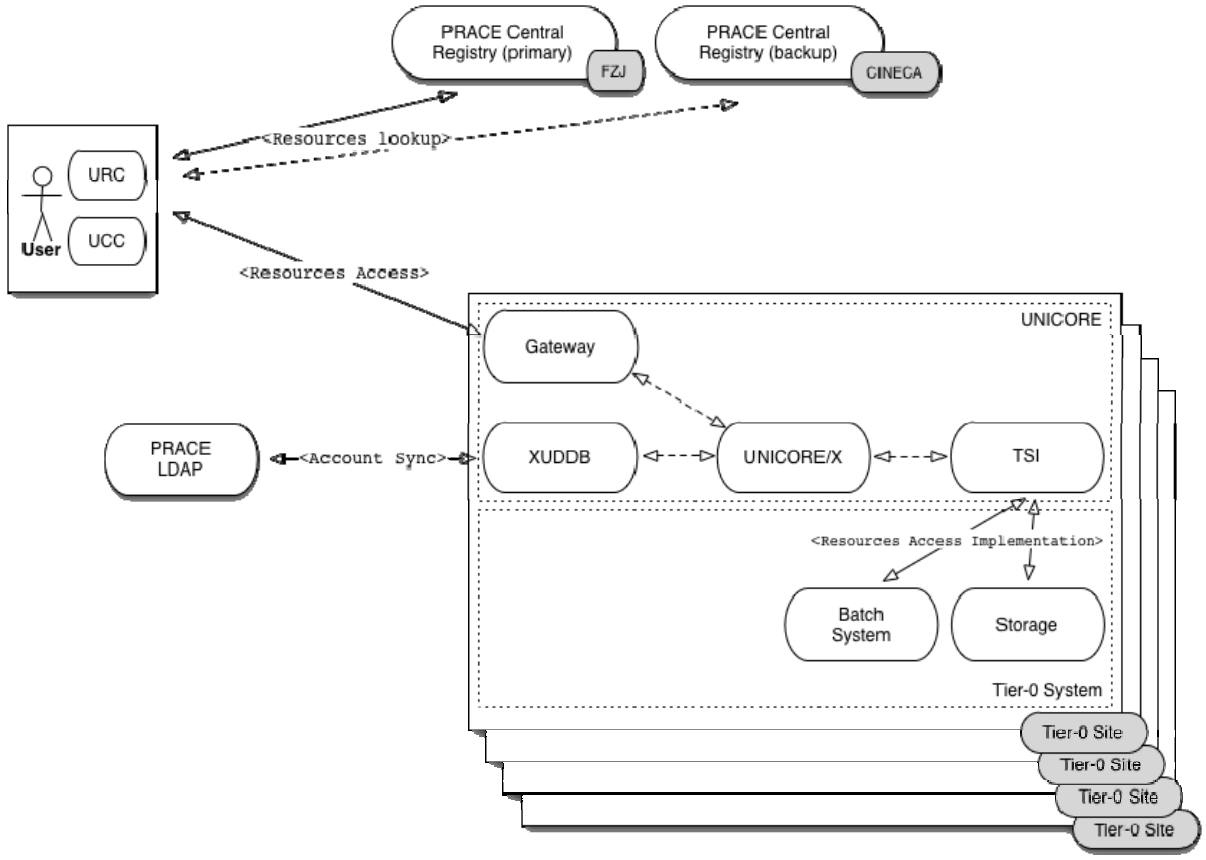

Figure 11: UNICORE deployment overview

The default version of UNICORE (6.4.2-p2) is currently installed on CURIE, HERMIT and SuperMUC while a newer version is already available on JUQUEEN.

On MareNostrum and FERMI it is planned to be available on July 2013.

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 [5] 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 [6], or by one of the two sister organizations TAGPMA and APGridPMA, all three federated in the IGTF [9]. These PMAs all require a minimum set of requirements for the CP/CPS (Certificate Policy/Certification Practice Statements) of the member CAs, as published and maintained in profile documents.

For PRACE, a distribution of CA information is maintained at a central repository. New IGTF distributions are made available there for the PRACE partners for download in several formats.

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 12: PRACE LDAP directory tree

. However, separate databases are used for Tier-0 and Tier-1 accounts. Information about Tier-0 accounts uses the “ou=ua,dc=prace-project,dc=eu” part of the name space. All six operational Tier-0 sites manage their own branch in the LDAP repository, three of which have been added this period.

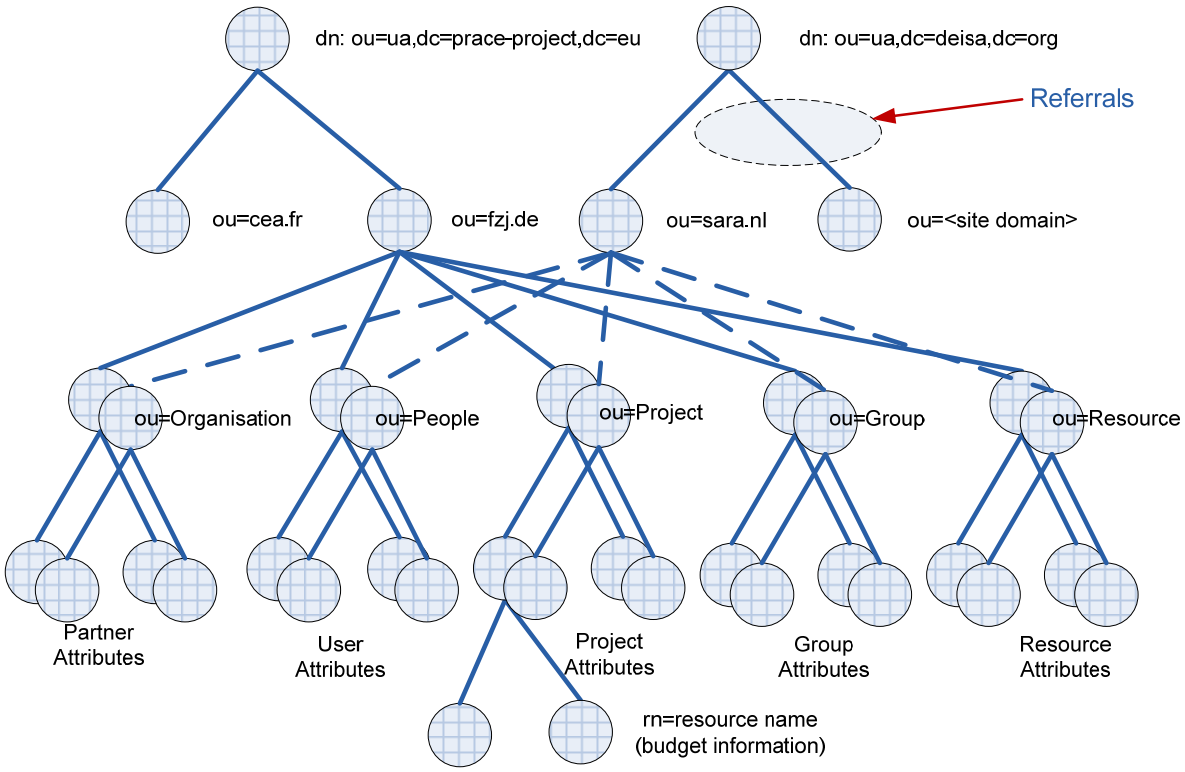

Figure 12: PRACE LDAP directory tree

The Tier-0 partners agreed in this period to adopt the same name conventions for the login names as used for the Tier-1 accounts. Name collisions are avoided by using this convention.

The coordination between the administrators of the projects that users submit with the PRACE Peer Review (PPR) facility and the Tier-0 partners has been improved. Tier-0 partners register the user accounts in LDAP but need information about the projects too. Depending on the choice of the Tier-0 partner either the PPR administrators add project information to LDAP or the Tier-0 partners add this information. In the latter case the PPR administrators provide the project information directly to the partner. It was also agreed among the Tier-0 partners to limit the number of LDAP attributes used for Tier-0 accounts to those needed. Some LDAP attributes used for Tier-1 users are not needed for Tier-0 accounts, basically because Tier-0 accounts are not shared among partners. For these reason these attributes are not used for Tier-0 accounts.

The Tier-0 partners accepted an efficient procedure for the provision of access for operational activities on their systems by PRACE staff from other partners. Accounts are registered through LDAP and each site publishes additional administrative requirements if needed in the PRACE internal wiki.

4.4.3 *Interactive access*

Interactive access to the Tier-0 systems is provided either by the SSH (Secure Shell) facilities, provided by most distributions of operating systems, or X.509 based SSH facilities. The latter are mostly used for access from other PRACE sites using the dedicated PRACE network. The most used implementation for the X.509 based SSH is GSI-OpenSSH [11] (GSI-SSH for short), distributed by the Globus community. GSI-SSH_Term, a GSI-SSH client, is supported by the PRACE partner LRZ. Details on how users can access their Tier-0 system are provided in the user documentation [12].

4.4.4 *Accounting services*

Users and PRACE staff can get accounting information in a uniform way with the DART tool [10]. With this tool users can display accounting information from all Tier-0 systems.

4.5 User Services

4.5.1 *PRACE Common Production Environment*

The aim of the PRACE Common Production Environment (PCPE) is to provide a working environment which has, at least as much as possible, a uniform interface from the user point of view.

The PCPE assumes that all Tier-0 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. Module names for each of the components of the PCPE are defined, although the PCPE module name can just be an alias to a module that already exists on the system. This flexibility using aliases for module names allows for the different system configurations at each of the Tier-0 sites. Site administrators can add additional tools/libraries to the PCPE at their site if they think they are beneficial to users. The approach taken regarding the implementation of the PCPE is to provide a uniform working environment to PRACE users without being overly prescriptive regarding the manner in which the system administrators achieve this.

One module called 'prace' is defined that enables access to the PCPE modules and sets the PRACE environment variables. The current set of modulefiles includes, see Table 6:

Tool	Description
bash	BASH Shell
tcsh	TC Shell
OpenSSH	Secure Shell
Emacs	Emacs Text Editor
nedit	Text Editor
C	C Compiler

Tool	Description
C++	C++ Compiler
Fortran	Fortran Compiler
Java	Java Compiler
Perl	PERL Programming Language
Python	Python Programming Language
TCL	Tool Command Language
TK	CL GUI Toolkit
gmake	GNU make
MPI	Message Passing Interface
BLACS	Basic Linear Algebra Communications
BLAS	Basic Linear Algebra Subroutines
LAPACK	Linear Algebra Package
ScaLAPACK	Scalar Linear Algebra Package
FFTW 2	Fast Fourier Transform in the West (v2)
FFTW 3	Fast Fourier Transform in the West (v3)
HDF5	Hierarchical Data Format
NetCDF	Network Common Data Format

Table 6: PCPE Component Modules

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.

To date the PCPE has been deployed on the Tier-0 systems as given in Table 7.

Site / Tier-0 System	PCPE Status
BSC / MareNostrum	Installed
CEA / CURIE	Installed
CINECA / FERMI	Installed
FZJ / JUQUEEN	Pending
HLRS / HERMIT	Installed
LRZ / SuperMUC	Installed

Table 7: Status of PCPE on Tier-0 Systems

Template modulefiles are available to download from the PRACE SVN repository which can be utilised by each of the sites. Examples for both linux and Cray XE platforms are currently available. Step-by-step guides are available on our internal PRACE Wiki to aid sites in the

configuration process. Again, the aim is to make the PCPE as flexible, but also as straightforward as possible to implement.

Work is continuing at present to fully integrate the PCPE in the PRACE Inca Monitoring framework. A set name has been defined for each core component that allows for a site-specific interface to the Inca monitoring framework to be implemented. This includes the ability to report on the different versions of each module which are deployed at each site. An example is illustrated by Figure 13.

Compilers	LRZ SUPERMUC
java version	1.6.0
Libraries	LRZ SUPERMUC
fftw version	3.3
hdf5 version	1.8.9
netcdf version	4.2.1.1
Shells	LRZ SUPERMUC
bash version	3.2.51(1)
tcsh version	6.15.00
Tools	LRZ SUPERMUC
emacs version	22.3.1
gmake version	3.81
nedit version	5.5
openssh version	5.1p1
perl version	5.10.0
python version	2.6.8
tcl version	8.5.5
tk version	8.5.5

Figure 13: Sample INCA Monitoring of PCPE

4.5.2 *User Documentation*

User Documentation for each of the PRACE services is available on the PRACE website [13].

Updates to documentation follow the same process as agreed in the PRACE-1IP project. All updates are subject to review by the PRACE User Documentation Panel. Documentation owners have been assigned for each of the key areas of the PRACE RI. The document owners are all part of the PRACE Documentation Review Panel. Any major changes or new documents must be reviewed by the panel before they will be posted online. Minor changes are routed directly to the Documentation lead for publication.

4.5.3 *PRACE Helpdesk*

Support for all of the Tier-0 and Tier-1 systems routes through the PRACE Trouble Ticket System (TTS). Queues are in place in the TTS for all of the current Tier-0 sites.

The primary interface to the Helpdesk is via a web interface (Figure 14). 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.

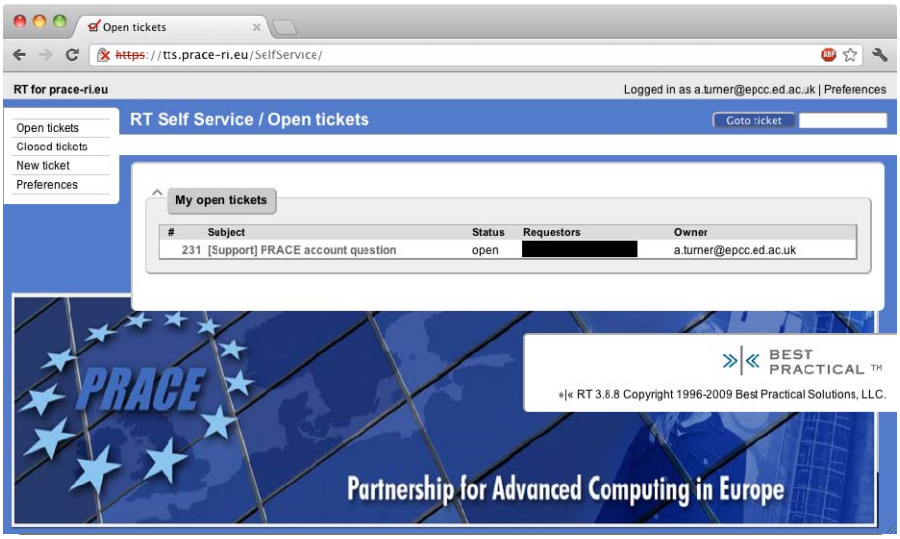

Figure 14: helpdesk web interface

Issues raised via the web interface are routed automatically to the site, thus minimising any delay in receiving support. A secondary email-based interface is also available. 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 Helpdesk team, who in turn route the issue to the appropriate site.

The TTS is used by both end users of the infrastructure, and internal PRACE support staff. Any infrastructure issues highlighted by the central PRACE monitoring services are logged by the Operator on Duty in the PRACE Helpdesk. This ensures that any issues are driven to conclusion in a timely manner.

4.6 Monitoring Services

Monitoring of services takes place using the Inca tool. 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. For this PRACE utilizes the user-level monitoring application Inca [14].

The efforts are a continuation of the efforts conducted for Tier-0 sites in PRACE-1IP, described in PRACE-1IP deliverable D6.3 [2]. The Inca configuration has been extended by the new Tier-0 resources and adopted where new services have been installed. For example new tests have been deployed to cover the version numbers of the middleware tools for remote access and data transfer, e.g. for the Globus Toolkit, GSISSH and GridFTP.

The new and updated resources comprise Globus middleware tests for CURIE@CEA, data transfer tests for HERMIT@HLRS as well as the Prace Common Production Environment (PCPE) test suite and the middleware tests for SuperMUC@LRZ. The resource JUGENE@FZJ has been removed as it went out of service. Furthermore preparations have been made for the integration of the new Tier-0 resources MareNostrum@BSC as well as JUQUEEN@FZJ.

5 Appendix A: PRACE Service Catalogue

This version is still under discussion, see section 2.2 for details.

The PRACE distributed research infrastructure provides a complete set of common services to its users. Service provision to users is done by the Tier-0 hosting partners, governed by the PRACE RI[1] 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 RI 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 Web, the PRACE RI 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-IP projects.

Tier-1 partners provide services to users as part of the DECI program for access, currently within the Implementation Phase projects.

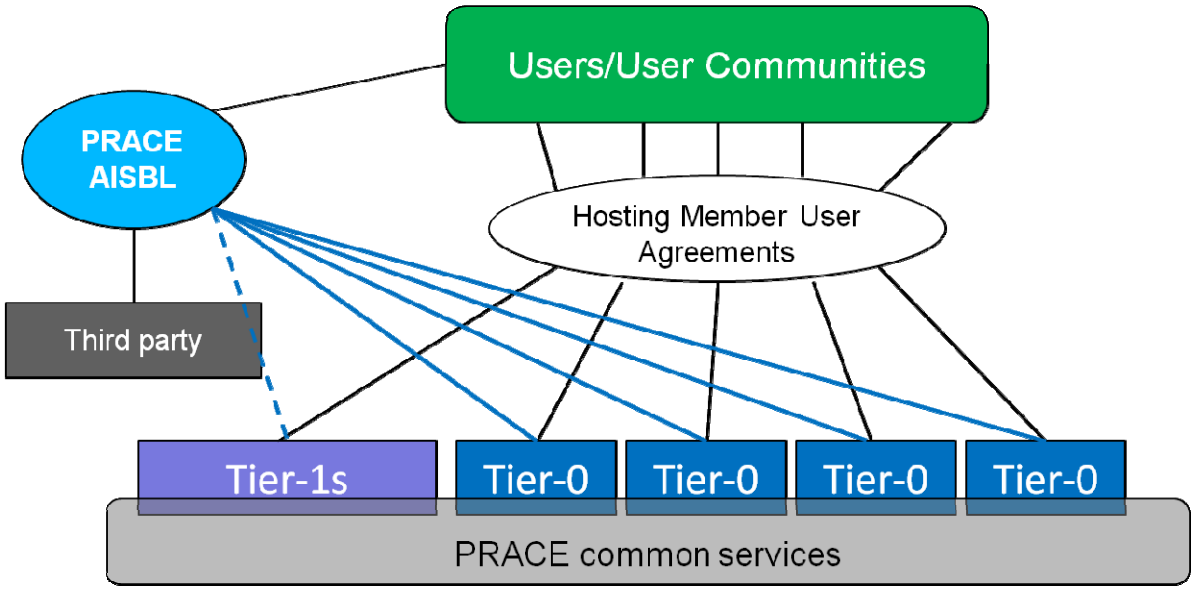

Figure 15: PRACE Service provision scheme and contracts to its users

To support a good and complete overview of all PRACE Operational Services, the PRACE Implementation projects have developed the PRACE Service Catalogue, which lists and describes the complete set of operational services that the PRACE RI 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 picture from hosting partners, other partners, the project and the PRACE RI.

[1] The PRACE RI is established as an international non-profit association (aisbl) with seat in Brussels

Classification of services

An important aspect of the PRACE Service Catalogue is the classification of services. Three service classes have been defined: Core services, Additional services and Optional services. The availability and support for each of these service classes is defined and described in Table 8.

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 RI 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 8: 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 RI or a single partner.

Update procedure

The PRACE Service Catalogue will be regularly updated to document the actual status of all services and will be maintained as a living document. 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.

New production services are proposed after an elaborate testing by evaluation teams, e.g. from the Implementation Projects. The evaluation is documented following the ISTP (Internal Specific Targeted Project) procedure. The acceptance as a production service is described by the Change Management procedure for PRACE operational services. The proposal for the new service also proposes the Service classification. In general a new service should be labeled additional or optional at the start. Later a service can be upgraded as core. After acceptance as a production service by the site representatives in PRACE operations this will be communicated to the Technical Boards of the IP projects and PRACE BoD with details of the new service. The new service will be added to a new version of the Service Catalogue. See Figure 16 for a schematic overview of the update procedure.

Twice a year the Service Catalogue will be reviewed in a dedicated video/telcon with all site representatives invited. After the meeting the updated Service Catalogue will be distributed for final comments to all site representatives and is accepted if no objections are received within 14 days. A new round of discussion is needed if changes are proposed, either by e-mail or video/telcon. For the updated version again the 14 day period for objections applies. This can be repeated until there is agreement.

After acceptance by the site representatives the new Service Catalogue will be distributed to the PRACE MB and the PRACE BoD for final acceptance.

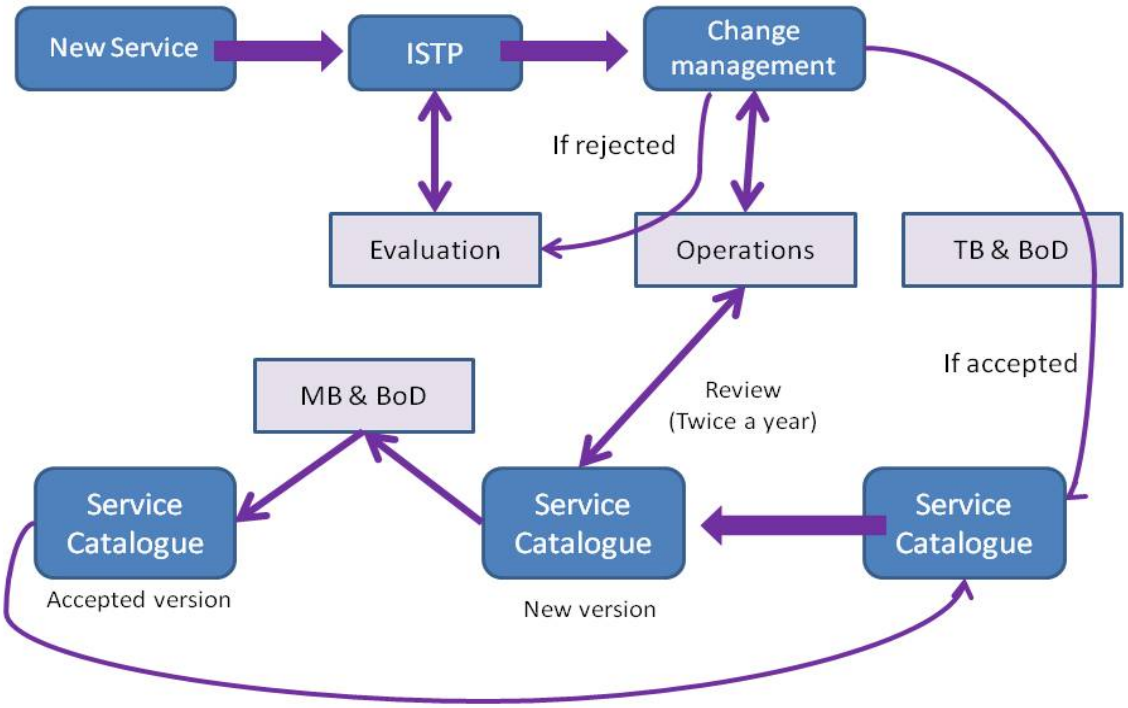

Figure 16: Update procedure Service Catalogue

PRACE Services

Table 9 at the end of this section gives a concise overview of all services and products.

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 3IP/2IP WP6 (compute services representative of the PRACE Operational Team)			
Reference:	Draft User Agreement			
Category:	Compute			
Service:	**Product/service**	**Tier-0**	**Tier-1**	**PRACE RI or single partner**
	UNICORE	core	core	-
	Globus GRAM	optional	optional	-
	Local batch system	core	core	-
Remarks:	The local batch system is included because it's a prerequisite for the provision of the other tools, however in general it will be different at each site.			

PRACE internal interactive command-line access to HPC				
Description:	Allows an employee or user of a PRACE partner to connect remotely to a Tier-0/1 system using the PRACE dedicated network and execute command-line instructions.			
Class:	Core			
Provider:	Tier-0/1 site + PRACE 3IP/2IP WP6 (compute services representative of the PRACE Operational Team)			
Reference:	NA			
Category:	AAA			
Service:	**Product/service**	**Tier-0**	**Tier-1**	**PRACE RI 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 3IP/2IP WP6 (compute services representative of the PRACE Operational Team)			
Reference:	Draft User Agreement			
Category:	AAA			
Service:	**Product/service**	**Tier-0**	**Tier-1**	**PRACE RI 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 a 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 RI 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 3IP/2IP WP6 (data services representative of the PRACE Operational Team)			
Reference:	Draft User Agreement			
Category:	Data			
Service:	**Product/service**	**Tier-0**	**Tier-1**	**PRACE RI or single partner**
	MC-GPFS	optional	optional	-
	GridFTP	core	core	-
	UNICORE (UFTP)	additional	additional	-
	Gtransfer	additional	additional	
Remarks:	GridFTP is a core service for Tier-1 only if a dedicated network is available Gtransfer details are in change #26: https://prace-wiki.fz-juelich.de/bin/view/PRACE/Operations/GtransferDeployment			

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 3IP WP4, PRACE 2IP WP4, Tier-0/1 site, PRACE Advanced Training Centres			
Reference:				
Category:	User			
Service:	**Product/service**	**Tier-0**	**Tier-1**	**PRACE RI 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 RI + PRACE 3IP WP6, WP7, WP3 + PRACE 2IP WP6, WP7, WP3			
Reference:				
Category:	User			
Service:	**Product/service**	**Tier-0**	**Tier-1**	**PRACE RI 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:	User

Data Visualization				
Service:	**Product/service**	**Tier-0**	**Tier-1**	**PRACE RI 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 the provision of credentials and identifying a user's X.509 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:	Tier-0/1 site + PRACE 3IP/2IP WP6 (AAA services representative of the PRACE Operational Team)			
Reference:	PRACE Security Policy and IGTF			
Category:	AAA			
Service:	**Product/service**	**Tier-0**	**Tier-1**	**PRACE RI or single partner**
	PKI	core	core	-
	MyProxy	additional	additional	core
Remarks:	My proxy is provided by multiple partners (e.g. as backup/disaster recovery). Each site must provide the Myproxy client functionality			

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 + DECI team + PRACE 3IP/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 RI or single partner**
	LDAP (user	core	core	-

Authorization				
	administration)			
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 + DECI Team + Tier-0/1 site + PRACE 3IP/2IP WP6 (AAA services representative of the PRACE Operational Team)			
Reference:				
Category:	AAA			
Service:	**Product/service**	**Tier-0**	**Tier-1**	**PRACE RI or single partner**
	Apache/CGI DART	core	core	-
	LDAP (user administration)	core	core	-
	GridSAFE common accounting repository			additional
Remarks:	GridSAFE details are in change #69: https://prace-wiki.fz-juelich.de/bin/view/PRACE/Operations/GridSAFEinProduction			

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 RI or single partner**
	TWiki	-	-	core
	SVN	-	-	core

Information Management				
	BSCW	-	-	core
	Prace-ri website	-	-	core
	e-mail lists			core
Remarks:	-			

Network Management				
Description:	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). PRACE partner's local network specialists and the PRACE NOC should support PRACE users in using the PRACE network infrastructure. The PRACE backbone will be dedicated, whereas local site connectivity of HPC systems and PRACE servers to the global Internet are public.			
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 RI or single partner**
	PRACE dedicated network	core	additional	-
	DNS (PRACE RI domain management)	-	-	core
	PerfSonar framework (Iperf)	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. Currently only Iperf is used for monitoring the network.			

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, available software and service availability.			
Class:	Core			
Provider:	Tier-0/1 site + PRACE 3IP/2IP WP6 (Monitoring services representative of the PRACE Operational Team)			
Reference:				
Category:	Monitoring			
Service:	**Product/service**	**Tier-0**	**Tier-1**	**PRACE RI or single partner**
	Inca	core	core	core
Remarks:	Inca for Tier-0/1 is for client support and PRACE-RI/single server for the server part			

Reporting				
Description:	Periodic reports of system utilization from the Tier-0/1 hosting partner to the PRACE RI.			
Class:	Core			
Provider:	PRACE RI + Tier-0/1 Hosting Partner			
Reference:				
Category:	Monitoring			
Service:	**Product/service**	**Tier-0**	**Tier-1**	**PRACE RI or single partner**
	Apache/CGI DART	core	core	-
	GridSAFE common accounting repository			additional
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

Software Management and Common Production Environment				
Provider:	Tier-0/1 site + PRACE 3IP/2IP WP6 + WP7			
Reference:	NA			
Category:	Generic			
Service:	**Product/service**	**Tier-0**	**Tier-1**	**PRACE RI or single partner**
	Modules	core	core	-
	PCPE	core	core	-
Remarks:	The module command can be implemented in different ways, e.g. by wrapper scripts			

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 3IP/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 RI or single partner**
	Helpdesk (RT-TTS)	-	-	core
Remarks:	For Tier-0 users the helpdesk of the Tier-0 site can be the first level support address			

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 3IP/2IP WP6 + WP7
Reference:	Draft User Agreement, PRACE 1IP D6.1
Category:	User

Advanced User Support				
Service:	**Product/service**	**Tier-0**	**Tier-1**	**PRACE RI or single partner**
	Helpdesk (RT-TTS)	-	-	core
Remarks:	The helpdesk is needed to support this service			

Service	Service class	Product	Tier-0	Tier-1	PRACE RI or single partner
Network management, Monitoring					
Dedicated network	core	PRACE Network	core	additional	
		DNS			core
Network management, Monitoring	core	PerfSonar framework (Iperf)	core	core	
Data					
Data transfer, storage & sharing	core	MC-GPFS	optional	optional	
		GridFTP	core	core	
		UNICORE	additional	additional	
		gtransfer	additional	additional	
Compute					
Uniform access to HPC	core	Local batch systems	core	core	
		UNICORE	core	core	
		Globus GRAM	optional	optional	
AAA					
Authentication	core	PKI	core	core	
Authentication	core	MyProxy	additional	additional	core
Authorization, Accounting	core	User Administration (LDAP)	core	core	
Accounting, Reporting	core	Apache/CGI DART	core	core	
		GridSAFE accounting repository			additional
PRACE internal interactive access	core	GSISSH	additional	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	core	core	
		PCPE	core	core	
First level user support, advanced user support	core	RT-TTS			core (tool)
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	core
Generic					
Information management	core	TWiki			core
		SVN			core
		BSCW			core
		prace-ri website			core
		e-mail list service			core

Table 9*:* Overview of PRACE services, categories and product classes