# SEVENTH FRAMEWORK PROGRAMME
# Research Infrastructures

## INFRA-2007-2.2.2.1 - Preparatory phase for 'Computer and Data Treatment' research infrastructures in the 2006 ESFRI Roadmap

# PRACE

# Partnership for Advanced Computing in Europe

**Grant Agreement Number: RI-211528**

# D4.2.2
# Deployment of enhanced solutions

## *Final*

Version: 1.0
Author(s): Riccardo Murri, ETHZ/CSCS
Date: 24.06.2009

## Project and Deliverable Information Sheet

| | | |
|---|---|---|
| **PRACE Project** | **Project Ref. №: RI-211528** | |
| | **Project Title: Partnership for Advanced Computing in Europe** | |
| | **Project Web Site:** http://www.prace-project.eu | |
| | **Deliverable ID:** < **D4.2.2**> | |
| | **Deliverable Nature:** <DOC_TYPE: Report / Other> | |
| | **Deliverable Level:**<br>PU * | **Contractual Date of Delivery:**<br>30 / 06 / 2009 |
| | | **Actual Date of Delivery:**<br>30 / 06 / 2009 |
| | **EC Project Officer: Maria Ramalho-Natario** | |

* - The dissemination levels 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:** <**Deployment of enhanced solutions** > | |
| | **ID:** <**D4.2.2**> | |
| | **Version:** <1.0> | **Status:** Final |
| | **Available at:** http://www.prace-project.eu | |
| | **Software Tool:** Microsoft Word 2003 | |
| | **File(s):** **D4.2.2.doc** | |
| **Authorship** | **Written by:** | Riccardo Murri (ETHZ/CSCS) |
| | **Contributors:** | Anton Frank, Jarno Laitinen (LRZ); Pekka Lehtovuori (CSC); Xavier Delaruelle (CEA); Miroslaw Kupczyk (PSNC); Bernd Schuller, Ralph Niederberger, Mathilde Romberg, Michael Rambadt, Achim Streit (FZJ); Vincent Ribaillier, Philippe Collinet (IDRIS); Michael Schliephake (HLRS); Gabriele Carteni (BSC) |
| | **Reviewed by:** | Jonathan Evans (BSC); Dietmar Erwin (FZJ) |
| | **Approved by:** | Technical Board |

## Document Status Sheet

| **Version** | **Date** | **Status** | **Comments** |
|---|---|---|---|
| 0.1 | 07/05/2009 | Draft | Laid out structure |
| 0.2 | 25/05/2009 | Draft | First full text draft |
| 0.3-0.7 | 09/06/2009 | Draft | Merged comments from WP4 mailing list and |

| | | | |
|---|---|---|---|
| | | | videoconferences. |
| 0.8 | 10/06/2009 | Draft | For PRACE internal review |
| 0.9 | 23/06/2009 | Draft | Merged PRACE reviewers’ comments |
| 1.0 | 24/06/2009 | Final version | |

## Document Keywords and Abstract

| | |
|---|---|
| **Keywords:** | PRACE, HPC, Research Infrastructure |
| **Abstract:** | The distributed systems management software stack already defined in D4.1.3 is further expanded and enhanced to provide better interoperability among the PRACE prototypes and existing European infrastructures like DEISA. |

**Copyright notices**

# Table of Contents

# References and Applicable Documents

[1] PRACE, http://www.prace-project.eu

[2] F. Berberich, E. Griffiths, “Report on options for a legal entity”, PRACE D2.1.1

[3] S. Requena, M. Kupczyk, “Operational model analysis and initial specification”, PRACE D2.6.1

[4] DEISA, http://www.deisa.eu/

[5] P. Kunszt, “Requirements Analysis for Tier-0 Systems Management”, PRACE D4.1.1

[6] X. Delaruelle, M. Kupczyk, R. Murri, J. Laitinen, V. Ribaillier, J. Bartolomé, “Evaluation report of existing solutions for ecosystem integration”, PRACE D4.2.1

[7] P. Kunszt, R. Murri, “Deployment of initial software stack to selected sites”, PRACE D4.1.3

[8] European Policy Management Authority for Grid Authentication, http://www.eugridpma.org/

[9] J. Reetz, Th. Sodderman, B. Heupers, J. Wolfrat, “Accounting Facilities in the European Supercomputing Grid DEISA”, GeS2007. Available online at: http://www.ges2007.de/fileadmin/papers/jreetz/GES_paper105.pdf

[10] DART, http://www.deisa.eu/usersupport/user-documentation/deisa-accounting-report-tool

[11] OGF UR-WG, “Usage Record Format Recommendation version 1.0”, https://forge.gridforum.org/projects/ur-wg, document id 15329

[12] Shibboleth, http://shibboleth.internet2.edu/

[13] W. E. Johnston, “The Evolution of Research and Education Networks and their Essential Role in Modern Science”, HPC2008 – High Performance and Grids, Cetraro, Italy

[14] PRACE WP6 “Top-10 HPC users” survey for D6.2.1 raw data, https://bscw.zam.kfa-juelich.de/bscw/bscw.cgi/d214395/PRACE_Top10-User_Survey.xls

[15] Globus Reliable File Transfer Service (RFT), http://globus.org/toolkit/docs/4.2/4.2.0/data/rft/index.html

[16] V. Welch, “Grid Security Infrastructure Message Specification”, OGF GFD-I.078, 2006. Available online at: http://www.ogf.org/documents/GFD.78.pdf

[17] S. Tuecke, V. Welch, D. Engert, L. Perlman, and M. Thompson, “RFC3820: Internet X.509 Public Key Infrastructure (PKI) Proxy Certificate Profile”, RFC3820, Internet Engineering Task Force, 2004. Available online at: http://www.ietf.org/rfc/rfc3820.txt

[18] R. Petrov, “OpenSSH secure shell and X.509 v3 certificates”, http://www.roumenpetrov.info/openssh/

[19] WP4 Wiki page on the “Enhanced Software Stack” technical details, https://twiki.cscs.ch/twiki/bin/view/PRACE/EnhancedSoftwareStack (Access is restricted, please register an account and write to riccardo.murri@cscs.ch to be granted permission.)

[20] SARA distribution of CA certs, http://winnetou.sara.nl/deisa/certs/

[21] eDEISA SA3, "Final report on GT4 (GRAM, GSI-SSH) and other middleware" Task T2b, eDEISA deliverable D-eSA3-B3

[22] DEISA2 WP4, "Initial report on technologies" DEISA2 deliverable D4.1

[23] DEISA2 WP4, "Progress on technologies in the first year" DEISA2 deliverable D4.2

[24] eXist, http://exist.sf.net/

[25] UNICORE, http://www.unicore.eu/

[26] CPMD, http://www.cpmd.org/

[27] DEISA Common Production Environment, http://www.deisa.eu/usersupport/primer/deisa-common-production-environment

[28] Iperf, http://iperf.sf.net/

[29] perfSONAR, http://www.perfsonar.net/

[30] IGTF, http://www.igtf.net/

[31] Globus, http://www.globus.org/

[32] The MCS Systems Administration Toolkit, http://www.mcs.anl.gov/hs/software/systems/msys/

[33] An Illustrated Guide to SSH Agent Forwarding, http://unixwiz.net/techtips/ssh-agent-forwarding.html

[34] OGSA BES, http://www.ogf.org/documents/GFD.108.pdf

[35] JSDL, http://www.gridforum.org/documents/GFD.56.pdf

[36] OGSA, http://www.gridforum.org/documents/GWD-I-E/GFD-I.030.pdf

[37] WP4 “module” configuration instructions, https://bscw.zam.kfa-juelich.de/bscw/bscw.cgi/295381

[38] SRB, http://www.sdsc.edu/srb/index.php/Main_Page

[39] iRODS, http://www.irods.org/

[40] WP4 evaluation report on the Sector/UDT software, https://twiki.cscs.ch/twiki/bin/view/PRACE/SectorTestbed

[41] Sector, http://sector.sf.net/

[42] Bandwidth Challenger winner at SC’08, http://sourceforge.net/forum/forum.php?forum_id=890270

[43] INCA, http://inca.sdsc.edu/drupal/

[44] Frank Zeller et al., "Deployment and operation of GridFTP and SRB", eDEISA deliverable D-eSA3-C2

[45] Frank Zeller et al. "Final report on SRB and other new data management technologies", eDEISA deliverable D-eSA3-C3

# List of Acronyms and Abbreviations

| | |
|---|---|
| AAA | Authorization, Authentication, Accounting. |
| API | Application Programming Interface |
| BES | Basic Execution Services; a standard Web Services-based interface for creating, monitoring, and controlling computational entities |
| BSC | Barcelona Supercomputing Centre, Spain |
| BSCW | "Be Smart – Cooperate Worldwide"; a web cooperation platform from OrbiTeam Software GmbH |
| CEA | French atomic energy commission, operating one of the supercomputing centres of GENCI |
| CA | Certification Authority |
| CGI | Common Gateway Interface |
| CINECA | Consorzio Interuniversitario, supercomputing centre in Italy. |
| CSC | IT Center for Science Ltd. Espoo, Finland. |
| CSCS | Swiss National Supercomputing Centre in Manno, Switzerland |
| CIS | Common Information System; a component of UNICORE |
| DART | DEISA Accounting and Reporting Tool |
| DEISA | Distributed European Infrastructure for Supercomputing Applications. EU project acronym. |
| DN | Distinguished Name |
| DUAS | DEISA User Administration System, a distributed LDAP system for user authorization. |
| ECMWF | European Centre for Medium-Range Weather Forecasts. |
| EGEE | Enabling Grids for E-sciencE; EU Grid project lead by CERN and successfully completed in 2004. Follow-ups are EGEE-II and EGEE-III. |
| EPCC | Edinburgh Parallel Computing Centre. |
| ESFRI | European Strategy Forum on Research Infrastructures; created roadmap for pan-European Research Infrastructure. |
| FTP | File Transfer Protocol |
| FZJ | Forschungszentrum Jülich Germany. Member of the German Gauss Supercomputing Centre. |
| 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. |
| GridFTP | Grid File Transfer Protocol, an extended version of the standard FTP protocol, making use of X.509 certificates for authentication. |
| GSI | Grid Security Infrastructure |
| HLRS | The High Performance Computing Center in Stuttgart, Germany. Member of the German Gauss Supercomputing Centre. |

| | |
|---|---|
| HPC | High Performance Computing; Computing at a high performance level at any given time; often used synonym with Supercomputing. |
| HTTP | HyperText Transport Protocol |
| HTTPS | HTTP over SSL |
| IDRIS | Supercomputing Center in Orsay, France. Member of GENCI. |
| IGTF | International Grid Trust Federation |
| INCA | User level grid monitoring tool used in DEISA. See references. |
| IP | Internet Protocol |
| iRODS | “I” Rule-Oriented Data System; a data grid software |
| JSDL | Job Submission Definition Language, an OGF standard. |
| LAN | Local Area Network |
| LDAP | Lightweight Directory Access Protocol |
| LRZ | Leibniz Supercomputing Centre in Garching, Germany. Member of the German Gauss Supercomputing Centre. |
| NREN | National Research and Education Network |
| OCSP | Online Certificate Status Protocol |
| OGF | Open Grid Forum, a standardization body for Grid Services |
| OGSA | Open Grid Services Architecture |
| PMA | Policy Management Authority |
| PKI | Public Key Infrastructure |
| POSIX | Portable Operating System Interface for Unix |
| PRACE | Partnership for Advanced Computing in Europe; EU project acronym. |
| PSNC | Supercomputing centre in Poznan, Poland. |
| RFT | Reliable File Transfer service by Globus, based on GridFTP. |
| SARA | The Dutch supercomputing centre in Amsterdam. |
| SRB | Storage Resource Broker; a data grid software. |
| SSH | Secure Shell. SSH is a network protocol that allows data to be exchanged using a secure channel between two networked devices. Used primarily on Linux and Unix based systems to access command-line shell accounts. |
| SSL | Secure Socket Layer |
| 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. |
| TCP | Transmission Control Protocol |
| UID | User IDentifier |
| UDP | User Datagram Protocol |
| UDT | UDP-based Data Transfer |
| UNICORE | Uniform Interface to Computing Resources. Grid software for access to distributed resources. |
| UR | Usage Record |

| | |
|---|---|
| X.509 | X.509 is an ITU-T standard for a public key infrastructure (PKI) and Privilege Management Infrastructure (PMI). X.509 specifies, amongst other things, standard formats for public key certificates, certificate revocation lists, attribute certificates, and a certification path validation algorithm. |
| XML | eXtensible Markup Language |
| XUUDB | UNICORE User DataBase; a component of UNICORE |

# Executive Summary

This document describes the enhanced software stack for the six PRACE prototypes selected previously as candidates of likely Petaflop/s systems in 2009/10. The enhanced software stack improves on the basic services implemented in the initial software stack specified in the earlier deliverable D4.1.3, and adds new services as required to fulfil the requirements set forth in PRACE deliverables D4.2.1 and D2.6.1 for ecosystem integration and operation of the PRACE infrastructure.

This document presents the server and software infrastructure set up by WP4 to support the enhanced software stack, along with the intended configuration of the PRACE prototype systems. Actual deployment of the software stack to PRACE prototypes will be performed by WP5.

The enhanced software stack is built upon widely-accepted standards, and interoperability with the DEISA infrastructure has been a primary objective: production DEISA software has been used where possible and adapted where necessary.

A new user administration infrastructure is specified, that can easily migrate user accounts from the DEISA Tier-1 federation to PRACE Tier-0 systems; the architecture can also accommodate different models of user administration (centralized or de-centralized).

In order to fulfil the vision of seamless access to the PRACE resources, an infrastructure based on UNICORE6 and several components of the Globus suite is set up. Command-line access through GSI-SSH and X.509-SSH is also supported, catering to the most common access pattern from current users.

A data movement infrastructure based on Globus RFT and the standard GridFTP is set up, to allow users to easily move large data sets to, from, and within the PRACE infrastructure in an unattended yet reliable way.

Finally, a prototypal PRACE monitoring infrastructure is specified, which WP4 expects to refine in later deliverables based on feedback from operations teams and sites. Access policies to the monitored information are discussed at length with their security implications.

# 1 Introduction

This document describes the enhanced software stack for the six PRACE prototypes previously selected as candidates of likely Petaflop/s systems in 2009/10. The enhanced software stack improves on the basic services implemented in the initial software stack [7], and adds new services as required to fulfil the requirements set forth in PRACE deliverables D4.2.1 [6] and D2.6.1 [3] for ecosystem integration and operation of the PRACE infrastructure.

The enhanced software stack is built upon widely-accepted standards, and interoperability with the DEISA infrastructure has been a primary objective: production DEISA software has been used where possible and adapted where necessary.

## 1.1 Background and previous work

Requirements for distributed systems management of prospective Tier-0 systems were collected and analyzed in the early WP4 deliverable D4.1.1 [5]. An important point made there is, that there is no foreseeable difference between systems management of (future) Tier-0 machines and (current) Tier-1 ones: solutions that are already in production use today in the HPC ecosystem are likely to be fit for tomorrow use as well. In particular, the systems software that is currently in production use within the DEISA [4] federation of national Tier-1 centres, is a suitable candidate to be the basis for managing the distributed PRACE infrastructure.

Ecosystem integration of PRACE systems was further investigated in D4.2.1 [6]: requirements were analyzed and candidate solutions were proposed to allow seamless access to PRACE resources and data exchange with national, regional and institutional computing centres and infrastructures. Adoption of standard protocols for data transfer and resource access was considered key to enabling ecosystem integration of the PRACE infrastructure.

An initial software stack was specified in PRACE deliverable D4.1.3 [7], covering the areas of User Administration and Accounting, Resource Management and Access, Distributed Data Management, and Monitoring of Distributed Resources. The initial software stack only covered basic functionality: the enhanced software stack described here improves and extends the services provided in order to fully implement the operational requirements coming from previous WP4 deliverables and the initial operational specification D2.6.1 [3].

## 1.2 Scope

This document presents the server and software infrastructure set up by WP4 to support the enhanced software stack, along with the intended configuration of the PRACE prototype systems. Candidate software, which had already been singled out in earlier deliverables [6] and [7], has been assessed (either directly by WP4 or relying on evaluations performed by DEISA) and weighed against the requirements stated in PRACE deliverables D4.2.1 [6] and D2.6.1 [3]. The final selection, together with a configuration suitable for the PRACE infrastructure, constitutes the object of this report; actual deployment of the software stack to PRACE prototypes will be performed by WP5. Volatile technical details of the software infrastructure (e.g., host names, TCP ports, firewall settings), have been collected into a Wiki page [19], where they will be kept up-to-date and consistent.

During this initial phase of the PRACE project, PRACE prototypes are operated by the sites hosting them. Since the operational model for PRACE during the initial phase has not yet been decided, we assume in this document that it will be initially close to the present situation, in which national sites operate the systems on behalf of PRACE (called "Cycles" model in [2]). Modifications suggested for future direct operation by PRACE (called "Operator" model in [2]) are described as possible future enhancements.

It is also foreseen that requirements coming from the industrial users of PRACE would require modification of some of the software stack or usage policies described here. Since this would only likely affect the operation of the Tier-0 systems, WP4 expects to address this in the final deliverable D4.1.4.

## 1.3 Document structure

The remainder of this document is divided into five sections, each one covering a different aspect of the distributed systems management: Prerequisites, User Administration and Accounting, Resource Management and Access, Distributed Data Management, and Monitoring of Distributed Resources. Within each section (with the exception of "Prerequisites"), the outcome of candidate software evaluations is briefly recapped, and then the configuration and installation of the selected software for the PRACE infrastructure is described.

A final chapter "Conclusions and future work" links the present activity with the planned developments of the PRACE distributed systems management stack (which will be finalized in the later deliverables D4.1.4 and D4.3).

# 2 Prerequisites

While it is expected that local operations teams have their own procedures and software to manage the network and the remote systems access, there are some requirements that local policies and operations procedures need to comply with for proper functioning of the system management stack.

## 2.1 Network links

Network connectivity is essential to integration in the computing ecosystem. The PRACE systems are connected with the rest of the European (and World-wide) HPC ecosystem by means of public Internet links provided by the GÉANT academic network provider.

Since PRACE prototype sites are also partners in the DEISA project [4], all current PRACE prototype sites are connected through the DEISA private link network (which also relies on GÉANT hardware). Therefore, it is expected that communication between PRACE systems at different sites can use the private network established by DEISA; it is also desirable (although not required) that interactions with the DEISA systems can happen over the private network as well, which can greatly enhance security and speed.

## 2.2 X.509 certification authorities

X.509 provides a means for secure authentication, and encryption services for secure communication and mutual trust between sites. In order to properly authenticate users and services, an X.509 PKI is required and was mandated in earlier PRACE deliverables D4.2.1 [6] and D4.1.3 [7]. To establish trust relations, the entire X.509 infrastructure relies on Certification Authorities that any involved party can trust to provide identity assertions.

For proper and correct operation of the authentication infrastructure, all sites in the ecosystem should accept the same subset of Certification Authorities. EUGridPMA [8] is the certifying body that currently accredits all European Certification Authorities used in Grid projects; the International Grid Trust Federation (IGTF, see [30]) unites EUGridPMA and similar authorities from America and Asia. Almost every person that has made use of Grid computing facilities in Europe (e.g., DEISA, EGEE) has a certificate issued by a IGTF authority.

It is required that PRACE systems accept all certificates signed by any CA accredited by EUGridPMA [8]. It is recommended that prototype sites install (and regularly update) the suite of CA root certificates accredited by the International Grid Trust Federation (IGTF, see [30]). Pointers to the files containing the CA certificates distribution have been made available by WP4 participants at [19].

# 3 User administration and accounting

The initial software stack proposed by WP4 in the earlier deliverable [7] provided an easy-to-implement but very restricted solution for user authentication and accounting, aimed at catering to the limited needs foreseen for the embryonic PRACE infrastructure. Here a fully-fledged user management and accounting solution is described, that can accommodate the requirements of a production HPC infrastructure.

## 3.1 User administration system

### 3.1.1 *Evaluation of candidate solutions*

Two software systems have been considered for deployment: the DEISA User Administration System [9] and Shibboleth [12]. Both have been already described as candidates in the earlier deliverable D4.1.3 [7].

***The DEISA User Administration System***

The DEISA project has defined and adopted an LDAP-based user management system [9]. Each participating site is assigned a unique UID range and a three-character prefix, and is free to publish accounts belonging to the assigned range in a local LDAP server, which is made accessible to other sites. A central LDAP server provides an aggregated view of the full infrastructure users' database (through LDAP referrals to the site servers); participating sites are expected to import the whole DEISA user list into the local user database. A custom LDAP schema is used to hold the UNIX account information, plus details of the projects that users belong to and flags indicating which systems/sites they should be given access to. Over the years, this has proven to be capable of managing hundreds of users coming from several different sites, yet leaving each site the full control over the local user database. In addition, LDAP is a secure scalable well-known technology and it is already in use all over the world to provide local authentication and user database services.

***Shibboleth***

Shibboleth [12] is considered a very promising technology and should definitely be considered for deployment on the future PRACE Tier-0 infrastructure. Some issues block its widespread adoption at present:

- Few countries have set up a national Shibboleth authentication infrastructure that covers all the Research and Education institutions.
- Shibboleth is still mostly web-oriented: many software services which are essential for access and usage of HPC systems are not ready for deployment in conjunction with Shibboleth authentication. This would require time and effort which are not available in the present set-up of PRACE.

### 3.1.2 *Adoption and deployment in PRACE*

It is expected that most users will initially apply for compute resources on PRACE systems only after having had their applications validated and run on national Tier-1 sites. Additionally, since most PRACE partners are also partners in DEISA, having a system in which user information can be easily migrated from DEISA to PRACE is a very desirable feature. In addition, DEISA has already established operational procedures for running its user management system; since PRACE Tier-0 sites will probably also be part of the DEISA infrastructure, using a software system and procedures that systems administrators are already familiar with, would lead to a faster and smoother set up of the PRACE production infrastructure.

PRACE will adopt the DEISA user administration system, with the following modifications:

- The LDAP tree holding information about the users must be rooted at a different DN than `dc=org,dc=deisa` – this should be done to keep the PRACE users separate from the DEISA users (the sets of partners in the two organizations do not coincide).

Individual sites are free to re-use the same LDAP server, with the provision that they serve the two sets of users in separate LDAP trees.

- The DEISA LDAP schema will be used to keep user data like in DEISA.
- No UID range or username pattern will be allocated to sites: systems administrators are free to assign UIDs and UNIX user names according to local practice. The local UID and user name allocated to a PRACE user might be different from one prototype system to another. It is however required that the certificate DN that is associated a user in the PRACE user LDAP is correctly mapped to that user's local UNIX user name/UID for all services that use X.509-based authentication/authorization.

An agreement on the usage of a common LDAP schema has been reached with the DEISA personnel at the last DEISA "All-hands" meeting (Munich, March 23-24, 2009). A revised LDAP schema was also presented during the same session. WP4 is currently collecting the information needed to populate the initial user database; after that, sites will be contacted to enable the local LDAP, and a central LDAP tree will be activated at SARA. Implementation details (e.g., LDAP root DN) and the LDAP schema description will be made available at the Wiki page [19].

***Modifications for different operation models***

The LDAP-based system can adapt to both a "central authority" and a "distributed authority" model. In the former case, local PRACE installations pull user definitions from the central LDAP, but the contents of the central LDAP are not imported from local ones, rather directly managed by PRACE personnel responsible for user account administration. The latter case is the one proposed above for adoption on PRACE prototype systems, and the model currently used by DEISA. A smooth transition from the "locally operated" to the "centrally operated" model is also feasible: as authority is transferred from local sites to a central management body, a site's LDAP data (and authority upon it) is moved from the locally operated server to the central one.

The proposed system also allows for quick adoption of a unified system of user accounts management between DEISA and PRACE.

## 3.2 Accounting

Compute resources are usually allocated to users in CPU-hours budgets; accounting records are the proof of usage of the allocated computing power. Due to the distributed nature of PRACE systems, the selected accounting system: must allow remote browsing of accounting records; must discriminate between different security levels (a user should not be able to see other users' usage, but site administrators should be allowed to inspect all records pertaining to their site); must allow exchange and comparison of usage records across machines of different type.

### 3.2.1 *Evaluation of candidate solutions*

The DEISA accounting system was evaluated and found to comply with all stated requirements. Therefore, no other solutions were considered.

***DEISA accounting system***

The DEISA project has developed its own accounting system, including a Java-based client, for gathering accounting data and providing a report to users: see [9] for an extensive description; some changes have been implemented in DEISA since it was published: the following text is concise but up-to-date. The DEISA accounting system is comprised of the following components:

- An *accounting data provider* reads the proprietary-format log files of site-local resource management systems, adds missing property values from the distributed user administration system (e.g., project name, user's X.509 subject DN), transforms these combined data into the Open Grid Forum "Usage Record" (UR) format [11] and stores it into a database. The original software distributed by DEISA uses the "eXist" database [24] as a backend, but this can be replaced by modifying the data provider program. Usage records are pushed to the database at a site-configurable frequency, but at least once per day.

- The *accounting information services* run at every site and allow authorized entities (e.g., users, project managers, site system administrators) to retrieve their role-specific subset of information. This information service is accessible through secure HTTP, access authorization is granted based on the requestor's X.509 certificate.

- The availability of accounting information in a standard format at every site makes it possible to uniformly process the data by a *report generator*. DEISA provides the Java-based program DART [10] for this purpose.

Indeed, users and systems administrators just need to run the DART program from the DEISA web site: DART collects the data from all the sites and presents a nicely formatted report; the accounting data is fetched using the users' X.509 credentials, so different classes of users have access to different information:

- Systems administrators can access all the usage records at their site.
- Projects' principal investigators can view records related to all activities being accredited to their projects budgets.
- Regular users can only see their own activity record.

### 3.2.2 *Adoption and deployment in PRACE*

PRACE will adopt the DEISA accounting system for its prototype machines: the DEISA project will make the HTTP distribution software and the DART graphical reporting tool available to PRACE sites and users. Two components must be installed locally at prototype sites:

- Each PRACE site should write its own data provider component to export the usage records information in OGF UR format, or re-use the one provided by DEISA.

- Each PRACE site should install an accounting information service to make the accounting information available over HTTP with appropriate X.509 authorization (which can be gotten from the LDAP-based system described in Section 3.1). Firewall rules should be set up, so that the HTTP collection endpoint for accounting data is accessible from the public Internet; sites should only rely on X.509 authorization for protecting data.

Since all PRACE prototype sites are also partners in DEISA, the above software infrastructure is already installed, and knowledge about its workings is available on-site; WP5 will provide the necessary modifications for the PRACE prototypes.

A minor, yet essential, modification is required to the DEISA DART tool: that accounting data is aggregated by certificate DN or by the "generic uid" provided in the LDAP user administration system, not by UNIX account UID (DEISA has unique UIDs across all its systems which will be no longer a feasible option for PRACE). Work to adapt the DEISA software for PRACE is underway at SARA (whose personnel authored the original DEISA system, see [9]). The PRACE-adapted DART tool will be finally made available for online launching (using the "Java Web Start" technology) from the PRACE web site.

### 3.2.3 *Future enhancements*

The accounting system architecture has shown some performance issues when processing queries spanning an extended time interval: these lead to a large data set being transferred to the client machine for processing, which can require a significant amount of time. Technical solutions are already being discussed within DEISA; it is expected that PRACE sites deploy them as soon as they are available and deployed within DEISA.

***Resource equivalence***

The DART tool must provide some conversion of CPU time spent on different systems, in order to show a unified report. The actual conversion rate has been determined for DEISA systems, based on a set of benchmarks. Similar conversion factors for PRACE Tier-0 systems (which would be embedded into the PRACE version of the DART tool) will need to be defined after the benchmark results have been collected by WP6.

# 4 Resource management and access

In the earlier deliverable [7], a thin software layer for accessing resources was described, targeting ease of implementation and deployment speed. An extended software stack for accessing the PRACE resources is described here, aiming at implementing the vision of a seamless and feature-rich access to PRACE systems, but still supporting the access methods that the HPC community is used to (as expressed in the WP6 survey [14]).

## 4.1 Grid access

According to D4.1.1 [5], usage of a standardized job description language and job submission interface are mandatory in order to fulfil the vision of users being able to submit to the whole Tier-0 infrastructure using a unified interface. The UNICORE 6 software suite was selected to provide Grid access satisfying this requirement.

### 4.1.1 *Evaluation of candidate solutions*

UNICORE [25] is a vertically integrated Grid middleware, which provides seamless, secure, and intuitive access to distributed resources and data. UNICORE allows users to submit and monitor single jobs as well as complex workflows. In its recent version UNICORE 6 is web-service enabled and OGSA-based. Several common open standards are implemented and supported to enable interoperability with other Grid technologies and e-infrastructures. Three clients exist to use UNICORE: the Eclipse-based UNICORE Rich Client (URC), the UNICORE command-line Client (UCC) and the High Level Programming API (HiLA).

UNICORE has been already used in the DEISA production system for several years now. Within DEISA WP4, seven sites have evaluated and tested UNICORE 6. The tests range from submission of simple job scripts to the submission of complex workflows at each site. The tests also included file transfer capabilities, in addition to performance and stress tests of server and client components. Important factors in the evaluation were the usability of clients, and quality of documentation and user support. Almost all test were successful (only the workflow component did not have the expected level of functionality, but developers are informed of this and a fix is being deployed); the selection of clients fit the requirement and preferences of the users; user support of UNICORE 6 was ranked “notable”: requests are answered normally within one hour and most bug fixes are available with one day.

The preliminary conclusion of this evaluation (the final test report is not available at the time of this writing; the DEISA experts have kindly shared their conclusions and experience with PRACE WP4) is that UNICORE 6 is a promising technology for distributed job submission, and that functionality and flexibility have increased significantly over UNICORE 5.

UNICORE’s standard interfaces (supports the OGF standards JSDL [35] and OGSA-BES [34]), together with its programmability through a high-level API, provide instrumentation for supporting access to the PRACE resources through community-specific tools (e.g., application portals tailored to the usage of a certain scientific community).

### 4.1.2 *Adoption and deployment in PRACE*

Based on the above evaluation, WP4 has decided to adopt UNICORE6 as the Grid technology of choice within PRACE.

UNICORE services can be divided in central services (workflow orchestrators, registry, CIS), and site-specific services (gateway, UNICORE/X, XUUDB). The UNICORE 6 central services accessible to PRACE users are installed at FZJ; the servers are shared with the DEISA UNICORE6 installation. Backup DEISA central servers are installed at CINECA and will be available to PRACE shortly. A table with IP endpoints of the UNICORE central services is kept up-to-date at [19].

All PRACE sites will install UNICORE version 6; it is currently installed (and available to PRACE users) at CEA on its WP5/WP7 and WP8 prototypes.

It is expected that WP5 will undertake installation of UNICORE6 on the prototype systems; complete installation instructions are available from the UNICORE site [25] and through DEISA. Prototype sites, being also part of the DEISA infrastructure, might choose to delay the installation and upgrade to UNICORE6 when they do it for the DEISA production infrastructure. At each site, one gateway service and one UNICORE/X per supercomputer should be installed; at least one XUUDB instance is also needed: this can be a shared instance (for the whole site), or one XUUDB per supercomputer. Assuming that a single firewall controls external traffic, and that no restrictions are in place in the traffic within the site’s LAN, just a single port (the gateway port) needs to be opened in the external firewall. This port is configurable; sites that are not using the default one should report this on the Wiki page [19].

## 4.2 Interactive command-line access

Interactive command-line access is considered a mandatory feature (see [5], and also [14] where it was requested by 100% of the surveyed users), but its impact on interoperability in the HPC ecosystem is limited: it is sufficient that clients exist for each protocol, that can be installed and supported at users’ sites.

### 4.2.1 *Evaluation of candidate solutions*

Two technical solutions are mature enough and compatible with the X.509 authentication infrastructure recommended in [5] and [6]; sites are free to implement either one, but at least one of the two access methods should be provided.

***GSI-SSH***

GSI-SSH provides interactive command-line access, supporting authentication through the Globus GSI mechanism (based on X.509 proxy certificates, see [16]-[17]). It is especially convenient for users, in that it implements "single sign-on": users create a proxy certificate once and it can be used to authenticate them against different Globus services for a user-chosen limited amount of time (by default, 12 hours).

GSI-SSH is already in use in all major HPC/Grid infrastructures worldwide: it is routinely used in the U.S. TeraGrid, has recently been classified as a core service in the DEISA infrastructure (after successful evaluation, upon which this report is based), and is deployed in the EGEE infrastructure on the nodes where direct login from users is possible.

Some PRACE sites have expressed security concerns about GSI-SSH: since the proxy certificate is stored into a file, any intruder, who can grab hold of that file, can then impersonate the user for any authenticated action (interactive login, file transfer) across the whole infrastructure.

However, the threat is mitigated by the limited validity of the proxy. In addition, all GSI-enabled services using a site-wide unique authentication database (called "grid-mapfile"), which maps authentication credentials (X.509 certificate subject DNs) to local UNIX account names: if a user proxy certificate has been stolen, all sites supporting GSI should remove the mapping, thus effectively banning the user from using GSI-enabled services.

#### *X.509-SSH*

An X.509-certificate based authentication module is available for the standard program OpenSSH [18]. Sites may choose to provide SSH access from the public internet using X.509-SSH instead of GSI-SSH.

It is possible to configure SSH (with or without the X.509 authentication module) for passwordless access to systems, using the SSH "agent forwarding" feature (see [33]). An X.509-SSH door node can thus be implemented: users log in to the door node, and can then log in to other systems, delegating authentication to the "agent forwarding".

When using "agent forwarding", the private key never leaves the user's client computer, so it cannot be stolen (like, for instance, X.509 proxy files). However, "agent forwarding" is vulnerable to a "hijacking" attack: any user that can communicate with the SSH agent can re-use its credentials to authenticate to a remote SSH server (see [33] for details). On the other hand, the "hijacking" attack can only be carried out while the user is logged in; short-lived sessions are less exposed to the risk. Automated logout after a certain amount of time could be implemented as a risk-mitigation measure.

X.509-SSH optionally offers the possibility of checking the certificate status with the Online Certificate Status Protocol (OCSP), which allows instant revocation of compromised certificates (as opposed to the usage of downloadable Certificate Revocation Lists, which are only updated every few hours). This would have the advantage that, as soon as the Certification Authority is informed of a certificate abuse or compromise, the certificate is immediately revoked and rendered invalid on all systems. On the other hand, not all CAs support OCSP currently, and the *online* verification introduces a dependency on the CA network servers.

Evaluation of X.509-SSH continues at HLRS.

### 4.2.2 *Adoption and deployment in PRACE*

No consensus was reached within WP4 partners regarding deployment of the SSH enhancements described above: differences in security policies could not be reconciled, to date, in a commonly agreed solution. WP4 expects to tackle the issue by promoting (together with WP2) a security forum where security officers from the different sites can meet and agree on a common acceptable solution. It is understood that the uniform access and "single sign-on" vision is a target for the PRACE production phase; while an agreement and a common policy is negotiated, it is acceptable that sites strengthen the access control policy to match the security level they consider acceptable.

GSI-SSH clients are quite widespread: they are available at most PRACE sites (through DEISA); they are also available through other computing projects (e.g., EGEE). X.509-clients are currently installed at HLRS.

Prototype sites are requested to enable either GSI-SSH or X.509-SSH access to the PRACE prototype. If no public Internet access is allowed, the prototype should be accessible at least via the private link network through a door node. This is a matter of local policy that WP5 will discuss with local personnel.

A "GSI-SSH door node" is available on the public internet (see [19]): this node acts as a gateway to access the rest of the PRACE infrastructure. PRACE users can log in to the door node, which has access to the PRACE private link network, and then log in to their target PRACE system (note that further steps are seamless, due to the "single sign on" feature of GSI-enabled services). The currently available GSI-SSH door node has been installed at LRZ; up-to-date information and IP endpoints are provided at [19].

Note: for the "door node" functionality to work, information about all PRACE users must be available to LRZ. This will only be operational when the centralized user administration system (described in Section 3.1) is in place. During the current transition phase, users should apply for an account at LRZ; up-to-date details are available at [19].

HLRS is currently considering whether to allow public Internet access via X.509-SSH, pending evaluation of this software.

## 4.3 PRACE user environment

Finally, users should be given a means of hiding inessential details such as software installation paths. The "module" framework has been selected for this purpose.

### 4.3.1 *Evaluation of candidate software*

WP4 relied on a earlier evaluation by DEISA, that compared `module` with the "SoftEnv" software [32]; results have been communicated by DEISA experts who also participate in WP4. The `module` framework was chosen and the validity of this choice has been proved by several years of production use.

### 4.3.2 *Adoption and deployment in PRACE*

In order to provide users with a uniform environment across several PRACE prototypes, the `module` command should be installed by WP5 on any PRACE prototype to control the default choice of compilers, tools, and applications, and configured as follows:

- The required version is the one provided by DEISA (link available at [19]).
- It is expected that users enter the command `module load prace`, before loading "module" settings for the individual applications supported by PRACE.

- The list of scientific application software provided by the `module avail` should be divided by scientific application domain. To ease transition from the DEISA infrastructure to PRACE, the categories should bear the same name used in DEISA; where possible, also the application invocation should be kept the same.

Work is currently underway to define a list of applications that should be provided on PRACE Tier-0 systems: a sample module file for the “CMPD” [26] application (which is both in the DEISA “Common Production Environment” and on the PRACE benchmarks list) is provided as an example. Its “module” configuration file will be installed by WP5 on all prototypes to enable the user-level application testing (see section 6.3). Details are available on the PRACE BSCW in the WP4 folder [37].

***Additional convenience scripts***

The script `prace_service`, available for download from [19], should be installed on all PRACE prototypes and made available to all users (possibly only after issuing “`module load prace`”). This command can provide an easy way of getting the endpoint associated with a specific Globus service at a particular site. The Globus services supported in PRACE are: GridFTP, RFT, GSI-SSH. For instance, the following command would open an interactive shell on the login node of a chosen PRACE site through GSI-SSH routed via the private link network:

```
sh$ gsissh `prace_service -i -s “site”`
```

The script itself is self-documenting, and will print out usage instructions if invoked improperly.

# 5 Distributed data management

It is a fact that scientific applications are used to process ever growing amounts of data, and this trend is not going to change in the next few years (see, for instance, [13]) actually, one should expect the data volume to grow even more as more data-intensive applications (such as the ones used by the bio-medical research community) are ported to HPC architectures. In the specific context of PRACE, the survey [14] run by WP6 to characterize the computational demands of the top 10 HPC users at PRACE partner sites showed (already in 2008) that HPC users would use PRACE systems to process data sets of Terabyte size, and many expressed the need to move data back to their home site or to the site where post-processing and visualization takes place.

There is thus a clear need for tools that can manage the movement of large data sets over high-speed networks: since transfer rates of large data sets can take hours or even days, such tools need to implement *unattended* and *reliable* management of data.

In addition, the selected tool must support transfer of data within the PRACE infrastructure (taking advantage of the high-speed private link) and movement to/from the rest of the HPC ecosystem over the public Internet. Client tools must be easily deployable at a generic site (user home site or pre/post-processing facility).

## 5.1 Evaluation of candidate solutions

### 5.1.1 *SIMDAT*

The SIMDAT software was originally developed as a proof-of-concept data grid, focused on providing the Meteorological community with services to share and discover meteorological data. WP4 has relied upon an evaluation provided by DEISA in the DEISA2 deliverables D4.1 [22] and D4.2 [23], which is summarized below.

A testbed has been deployed across three DEISA sites EPCC, CINECA, and ECMWF, in order to find out whether the software could be adapted for general scientific use. Tests were carried out in parallel by all participating sites to discover local and external datasets and to retrieve the discovered datasets. Initial results uncovered some weaknesses of the software, namely:

- Catalogue synchronization problems: Some deficiencies in the documentation regarding naming conventions for the data repositories caused the metadata of some datasets to not synchronize fully across to all sites. These problems were solved by changing the software configuration at all sites. The changes made where notified to the SIMDAT developers to be included in the next version of the documentation.
- Dataset download issues: Retrieval of datasets was not working when the connection between two nodes was established via a third party request. The solution implemented was to configure the testbed as a full-mesh network where all nodes can make the retrievals directly from the nodes hosting the data. A possible enhancement of the software to dynamically support third party routing was noted and the SIMDAT developers were informed.

DEISA decided not to adopt the SIMDAT software for production use. It was decided to discard SIMDAT in PRACE as well.

### 5.1.2 *SRB*

SRB [38] is a very high-level data management tool; according to its web site, “SRB supports shared collections that can be distributed across multiple organizations and heterogeneous storage systems. *[...]*SRB presents the user with a single file hierarchy for data distributed across multiple storage systems. It has features to support the management, collaboration, controlled sharing, publication, replication, transfer, and preservation of distributed data.”. WP4 relied on an evaluation provided by the DEISA project in eDEISA deliverables D-eSA3-C2 [44] and D-eSA3-C3 [45], which is summarized here.

The main assessment criterion was to verify whether the tool is reliable, secure and stable and if it can manage huge data and data facilities.

Test SRB servers were deployed at CINECA, CSC, HLRS, IDRIS and LRZ. These servers were interfaced to the DEISA private link network, but whenever tests required it, they were also connected to the public Internet.

A collection of SRB systems that share the same meta-data catalogue server (MCAT) is called a “SRB zone”. The DEISA testbed was comprised of several federated zones: users would work primarily in their site-local zone, but occasionally access the SRB server at a different site to browse collections, query metadata and access files they have permission to read. This is the model apt for the DEISA Tier-1 federation, but also for deployment into PRACE during the initial phase (and in the “Cycles” operational model).

The evaluation was completed and SRB was deemed a stable and mature software, complete with GridFTP and GSI integration. However, permissions on the group level were found too coarse grained (they can only be granted on a domain basis). Moreover, further SRB development has stopped in favour of iRODS. No production use is foreseen for SRB in DEISA.

It was thus decided to discard SRB in favour of other solutions also in PRACE.

### 5.1.3 *iRODS*

iRODS [39] is the successor to SRB. According to its web site: “The iRODS system is based on expertise gained through nearly a decade of applying the SRB technology in support of Data Grids, Digital Libraries, Persistent Archives, and Real-time Data Systems. iRODS management policies (sets of assertions these communities make about their digital collections) are characterized in iRODS Rules and state information. At the iRODS core, a Rule Engine interprets the Rules to decide how the system is to respond to various requests and conditions. iRODS is open source.” Essential parts of the iRODS architecture are a database management system for storing and querying persistent information and a “Rule Engine” that invokes micro-service workflows. iRODS features include:

- maintaining a *logical namespace* for identifying objects such files, users, storage systems, rules, micro-services, and persistent state information,
- data sharing based on flexible access control mechanisms,
- uniform access to distributed data stored in heterogeneous storage systems,
- storing of system and user-defined metadata,
- discovery of data based on queries,
- efficient data transport optimized for both small and large files (through parallel I/O in the latter case),
- management of data distribution and replication.

An evaluation of iRODS was conducted by DEISA (see eDEISA deliverable D-eSA3-C3), to verify if iRODS and its novel components are reliable, secure and stable. A testbed was deployed at FZJ, IDRIS and HLRS; the servers were connected to the DEISA private link network and configured for using GSI authentication.

Although iRODS has finally been deemed quite stable and reliable, DEISA experts reported to WP4 that iRODS requires the installation of some third-party software, which makes the installation procedure more complex. Moreover, they were not able to successfully exploit the GSI authentication, and had to fall back to username/password combination. Finally, the DEISA evaluation concluded that there is no real use case that makes it needed in the current production DEISA infrastructure.

Therefore, although iRODS seems to be a promising software and could be re-evaluated at a later stage of PRACE (possibly in conjunction with the DEISA production deployment, if it ever happens), it is deemed not viable for use in the current infrastructure.

### 5.1.4 *Sector/UDT*

According to its web site [41]: “Sector supports distributed data storage, distribution, and processing over large clusters of commodity computers. Sector is a distributed storage system.” The Sector software won the bandwidth challenge at the “SuperComputing” conferences SC06 and SC08 [42], and is then an interesting candidate software for high-speed data transfer.

A testbed was set up at CSCS and PSNC to evaluate Sector; versions 1.19 and 1.20 of the software were tested. The Sector system is comprised of 4 kinds of distinct servers:

- one *master* server, dispatching client requests to slave nodes
- one *security* server, providing authorization information

- any number of *slave* servers, providing the actual storage space.

The master and security servers were originally installed at CSCS together with one slave/storage server, with PSNC providing an additional slave/storage server; later on, the deployment was reversed and the master and security servers were moved to PSNC.

Several shortcomings were found in the Sector system (see details at [40]), mostly due to the relative immaturity of the code base:

- The master server is quite fragile, and crashes upon receiving an unexpected network packet; this includes network probe packets, therefore giving each user the possibility to crash the whole infrastructure (albeit unwillingly).
- When the master server crashes, all slave servers must be restarted in order to reconnect; currently the only supported way to do this is to allow passwordless SSH access from the master node to the slave nodes, which is unacceptable for security reasons.
- The client tools hard-code the location of the configuration file. Client authentication credentials are stored in configuration files in plain text, and there is no support for X.509-based authentication.
- All Sector servers authenticate basing on a X.509 certificate *file*, instead of using the certificate subject DN. Therefore, each time a component certificate changes (e.g., the certificate expires), all other infrastructure components must deploy the new certificate file at the same time. This poses quite a lot of coordination and operational problems on a production distributed infrastructure.

The obvious conclusion is that Sector is not yet mature for production use

### 5.1.5 *Globus RFT*

The Globus RFT service [15] is already in production use in eDEISA and proposed in DEISA2 (see [21], [22], [23]). It is a “central” service: only one server needs to be accessible (within a given infrastructure) to drive the transfers. Users can schedule data transfers between two GridFTP endpoints with the RFT server; the server will execute the request automatically and retry in case of failure.

RFT clients, like most Globus software, are already available in popular Linux distributions; compilation from source is also possible with few hassles.

Based on earlier evaluations by DEISA and on positive feedback by WP4 experts also participating in DEISA operations, WP4 has chosen RFT for deployment on the PRACE prototype infrastructure.

## 5.2 Adoption and deployment in PRACE

The Globus *Reliable File Transfer* (RFT) [15] software satisfies the requirements stated in Section 5.1: users request file transfers to a central server, which executes the requested transfer over the GridFTP protocol, without any further intervention. The transfer is retried if failed, ensuring that the file eventually reaches destination.

It was already recommended in [7] that all PRACE systems are accessible through GridFTP, but only a basic set of GridFTP command-line clients were recommended for installation. A new service layer is built on top of this, which can cater to the above requirements and improve users’ experience of the PRACE infrastructure:

- A central RFT server is installed and made accessible to all users of PRACE, which can be used to move data both within the private link network used by PRACE systems *and* within the public internet. The PRACE RFT server (currently shared with DEISA) is installed at LRZ (connection details are available at [19]).
- The command-line RFT client must be installed at all PRACE sites (they already are at the sites that participate in DEISA), so that users can schedule data movement among Tier-0 systems and their home or post-processing sites.
- The GridFTP endpoints that are used by RFT and other command-line clients should actually move data to/from a file system that is directly accessed by the PRACE systems front-end nodes. Where this is not possible by site restrictions, then it is the responsibility of the site to specify the procedure required to move data from the GridFTP-accessible storage area to the HPC machine file system, and make it available to PRACE users.

The recommended Globus version for RFT is 4.0.x, for compatibility with DEISA. Sites are expected to install version 4.2 when DEISA switches to it.

Based on security restrictions and network topology, we can divide the data movement instances into the following cases.

### 5.2.1 *Data movement between PRACE systems*

Data movement within the PRACE infrastructure can take advantage of the dedicated network links, and the trust relations that exist between PRACE sites. It is expected that GridFTP, when used between two PRACE systems, allows users to move data from one system to the other and have it readily available for use at the destination site; that is, GridFTP endpoints should connect file systems that are directly accessed by the PRACE systems.

Data movement happening as part of a workflow execution is dealt with by the workflow software directly (indeed, UNICORE provides an integrated data transfer interface that is available to PRACE users), so the above configuration is recommended as a convenience for users to quickly provide input files to applications or retrieve results.

No global distributed file system is mandated at this stage between PRACE systems (see also [5], [3]): all currently available solutions seem to present some shortcoming that makes them undesirable. An exhaustive technical evaluation of major distributed file system products is currently being performed by DEISA, but results will only be available in 2010. It should also be noted that the DEISA Tier-1 infrastructure substantially differs from PRACE in that DEISA started with a few machines with a very similar architecture, so the idea came naturally that users could easily change machines and therefore would appreciate a common global file system. Within PRACE, it is expected that machines would be very different in architecture, so the advantages of having a global shared file system are less relevant.

The RFT/GridFTP based data movement can actually provide most of the features of a distributed file system, with the exception of real-time data sharing.

### 5.2.2 *Data exchange with other sites/infrastructures*

A public GridFTP endpoint must be provided on the public internet in order to exchange data with other sites and infrastructures. For security reasons, not all sites might want to open their firewall to public GridFTP traffic. The DEISA project has successfully used “GridFTP door nodes” in the production infrastructure, analogous to “bastion hosts” used for SSH: one site enables GridFTP transfers to a “transit” storage area, and GridFTP over the private network link is used from there to the final destination system.

A “GridFTP door node” on the public internet is provided by one PRACE site, to which any PRACE user can connect to upload/download files into a temporary staging area. The GridFTP door node at LRZ is available to PRACE users; up-to-date connection details are available at [19].

The PRACE RFT server can be used to schedule transfers between any GridFTP endpoint on the Internet and the public “GridFTP door node”. In addition, the PRACE RFT server can be used to perform transfers from the temporary staging area and any PRACE system.

As already noted in previous PRACE deliverables (see [6]-[7]), GridFTP is currently the only *lingua franca* of data transfer. The public GridFTP door node provides all the functionality needed for exchanging data with other infrastructures (e.g., EGEE and U.S. TeraGrid, both of which use GridFTP-based data transfers) or users’ home sites.

### 5.2.3 *Data movement with UNICORE*

UNICORE (see section 4.1) offers a storage management service that allows file upload and download, and also server-server transfers that only need to be triggered by a client. This offers an alternative interface and route for data movement. A variety of protocols are supported by UNICORE, the default being HTTPS through the secure UNICORE channel (a single port in the firewall needs to be opened, see section 4.1).

For data upload/download from/to a client and server-server transfers, currently HTTPS (fast, but only bulk data) and OGSA-ByteIO (quite slow but with a rich, POSIX-like interface) are supported. For data staging (i.e. data movement prior to job submission and after the job ends) currently HTTPS, FTP, UDT and GridFTP are supported.

Many instances of the storage types are typically provided: job working directories, user's home directories, and dedicated file systems. Access to the *home* directory and each job’s working directory is automatically enabled in UNICORE6. Every other storage location should be defined in the UNICORE/X component by systems administrators. It is recommended that PRACE prototype sites provide at least access to a staging location from whence data can be moved to any other location on the site (by means of locally-provided tools and procedures).

# 6 Monitoring of distributed resources

Provision of up-to-date information about system and software status is crucial to the operations of an infrastructure by system administrators and to its productive usage by end-users. The following specification of a monitoring infrastructure is an initial proposal. WP4 expects to refine it and add or remove functionalities based on feedback by WP5 and the sites’ operations teams. It is also expected that the final choice of an operational model will have an impact in the long run.

## 6.1 Access to monitored information

Display of system status information is very much subject to security-sensitive issues; thus, stakeholders of monitored information must be clearly defined. Some information will only be of interest to a specific category of stakeholders (e.g., systems administrators); some is of general interest and can be published with little restrictions (e.g., applications available on a certain system); some should be restricted because it has potentially a security or privacy impact (e.g., details on a running job).

### 6.1.1 *Methodology*

The content of this chapter on monitoring is based on the outcome of a survey, which was sent to PRACE site representatives. The goal of this survey was to provide a matrix of resources to be monitored versus groups/roles of the people involved in PRACE usage and operations. The outcome of the survey has formed the basis of WP4 discussions that prepared the policy and technical implementation that is described here.

### 6.1.2 *Role-based access*

The following roles (each of which corresponds to a certain class of stakeholders) form the basis for access restrictions on monitoring information:

- General public: this information can be made available to everyone that has access to a PRACE system (possibly, everyone that is in possession of a valid X.509 certificate from a trusted CA).
- PRACE users of a site: this category comprises all users that have access to a specific PRACE system.
- Principal Investigators of a project.
- PRACE systems administrators: this category comprises all systems administrators across the whole PRACE infrastructure, and also PRACE personnel that is working on user support (“help desk”, which is currently being discussed in WP2)
- Local system administrators: this category includes systems administrators of a specific PRACE system.

### 6.1.3 *Modifications for different operational models*

In an operational scenario where the local sites operate the PRACE systems, it is expected that each site does its own monitoring, and that systems administrators of other sites only need enough information as necessary to pinpoint which site should take responsibility of a problem. Basically, the only information that is needed (infrastructure wide) is the network status and the status of the service endpoints at each site.

In an operational scenario where PRACE personnel directly runs the PRACE systems, systems administrators will have access to the full spectrum of information provided by low-level and local systems monitoring software, and it is not necessary to establish a policy for information exchange.

## 6.2 Network monitoring

Network performance is critical for any kind of application and can vary over time, depending on network parameters, routing, and parallel network load.

### 6.2.1 *Evaluation of candidate solutions*

DEISA installed a monitoring tool which tests network throughput values periodically. A small application has been programmed, which collects network throughput between all

DEISA sites three times a day. At every location `iperf` [28] TCP and UDP servers have been installed to which all DEISA sites (one node of every DEISA supercomputer) connect and measure performance data. The information available is constantly updated and provides the network administrators with a graphical overview showing daily, weekly, and monthly network throughput graphs between clients and servers selected as well as real logging data if needed for analysis.

Network monitoring is mostly used by systems administrators to diagnose infrastructure and communication problems, so it is deemed important that PRACE systems use an interface and a monitoring infrastructure that its systems administrators are already familiar with. As all PRACE prospective Tier-0 sites are also participating in DEISA, it is convenient that the monitoring system is actually the same.

### 6.2.2 *Adoption and deployment in PRACE*

The network monitoring infrastructure does not change from the one described in [7]: `iperf`-based network tests are run from any prototype site within PRACE to any other, and displayed in a matrix form.

The network monitoring pages are already available to prototype systems administrators (the PRACE site matrix is a submatrix of the DEISA one).

The monitoring software currently recommended for use in PRACE and DEISA also displays IP addresses of endpoints. Since users are interested in service availability more than in the details of IP networking, PRACE WP4 recommends that these IP addresses are replaced by names of the form “site-service” (e.g., “FZJ-GridFTP”) before making the network monitoring pages available to PRACE users. Since those DEISA web pages are mainly CGI-based currently, there is no technical problem in providing alternate pages, which hide detailed information to users. This modification is currently being worked out at FZJ, and will be rolled out when the network monitoring pages are made available to general users.

### 6.2.3 *Future enhancements*

DEISA is currently evaluating a new network monitoring system, based on the “perfSONAR multi-domain monitoring” suite [29]. This software suite provides standardized access to network monitoring information: interface statistics of routers, `iperf` throughput measurements, one way delays, light path circuit information, etc. Through this standardized interface network administrators, system administrators and users can gain access to the measurement data allowing to discriminate which user may access which level of information.

PRACE systems and network administrators must be able to monitor the status of all network paths involved in PRACE operations, which is required for debugging and alerting on network-related problems. Many NRENs already make such network fabric monitoring information publicly available via applications set on top of these perfSONAR services. It will be desirable to provide similar procedures and interfaces in DEISA and in PRACE. If the DEISA evaluation is positive, PRACE will switch to using the new system when DEISA does.

## 6.3 Software version monitoring

### 6.3.1 *Evaluation of candidate solutions*

DEISA is using the INCA [43] framework for monitoring installed software applications, and displaying a summary page.

The INCA system works as follows. INCA *reporters* run (as a normal, unprivileged user) on each local resource, retrieve version information and send it back to a central collector. INCA supports X.509-based authentication to protect sensitive data; user with different roles can be given authorization to access a different set of pages.

This INCA set up has been in DEISA production usage already and is thus considered adequate for adoption into PRACE.

### 6.3.2 *Adoption and deployment in PRACE*

Software version monitoring will be implemented in PRACE using the same INCA framework that is already in use within DEISA. Users coming from Tier-1 systems will find the same report style they are already accustomed to.

An INCA server, installed at LRZ, aggregates the results and displays them at http://inca.prace-project.eu. PRACE users will be given the rights to view these data, based on X.509-certificate authentication. Implementation of this policy requires the full list of PRACE users and their certificate subject DNs, so it will actually be rolled out when that is available (see section 3.1).

All applications defined in the “PRACE user environment” list will be monitored: for each one, the default version is reported on the web summary page. However, complete application monitoring is a goal for the final implementation phase; during the prototype phase, only availability of “CPMD” (through invocation of “`module load cpmd`”) will be monitored as an example on the current PRACE prototypes.

Information about installed software and OS will be made available to all PRACE users; there is little to be gained in restricting access as the information would likely be stated in publicly-available documentation (web site, training material); however, details that can be security-sensitive (sub-version of the Linux kernel; system libraries release; version of the command-line access software and protocol) should remain hidden from the general users, to prevent easy access to information that could be exploited by attackers.

## 6.4 Job monitoring

There are two separate interfaces for job submission available on the PRACE systems: UNICORE and the local batch system commands available from the command-line. There is no expected use case that requires users to mix the two submission interfaces and jointly monitor job progress. Therefore, it is deemed sufficient that jobs submitted through UNICORE are monitored through the UNICORE monitoring interface, and that jobs submitted through the local batch system commands are monitored through the corresponding locally available facilities.

It is expected that every user that can log in to a PRACE system has access to aggregate information on batch system status, such as: number of free job slots, number of running and queued jobs, estimated start time of jobs, etc. There are standard commands providing this in any modern batch system, so no problem is foreseen here.

## 6.5 Storage system status

Full information on storage systems should only be made available (and is only useful) to local systems administrators. PRACE users should have access to their own storage quota information (e.g., how much space has been used in the reserved quota; how much free space is left); where technically feasible, projects' Principal Investigators should be able to monitor the project's storage area.

As storage systems are entirely managed by local sites, it is up to the sites to define their own monitoring metrics and software for storage.

## 6.6 System and service status monitoring

### 6.6.1 *Evaluation of candidate solutions*

DEISA is using the INCA [43] framework for monitoring service availability and status, and displaying a summary page, as already described in section 6.3.

This INCA set up has been in DEISA production usage already and is thus considered adequate for adoption into PRACE.

### 6.6.2 *Adoption and deployment in PRACE*

Service status monitoring will be provided for PRACE prototypes using the same INCA framework described in Section 6.3. By construction, the INCA framework can only display information that reporters running at local sites send back to central server. It is recommended that sites enable at least the following INCA reporters to run and probe the prototype system:

- Validity of X.509 certificate (for all services that use X.509 host authentication); the regular users' view should only tell if a certificate is valid; systems administrators should be able to see what exactly is failing in the validation chain.
- Data transfer endpoints and servers: this should provide a view on the functional availability of each GridFTP endpoint and the RFT server.
- Remote job submission services: users should be presented with a status page of the UNICORE servers.
- Availability of the remote command-line access servers (GSI-SSH, X.509-SSH).

Separate views over the service functionality tests will be provided (using X.509-authorization to discriminate among roles):

- A minimal view should for all PRACE users, from which one can determine whether a certain service (e.g., a GridFTP endpoint) is available and operational at a site.
- A general view of the computational resources availability: up/down, free number of cores, number of jobs (running, pending), total quantity of particular resources. It is agreed that current quantity of particular resources can be published on the public website.
- A detailed view for local systems administrators and user support personnel, that includes information of why a certain service failed the availability test (e.g., GridFTP not functional because the server certificate has expired). It is not *currently* expected that PRACE systems administrators of other sites have access to this information (see section "Modifications for different operational models"), but this policy could be revised if the "operator" model is adopted.

Work on implementing the above views is still underway: interaction with site security officers and local systems administrators is needed.

### 6.6.3 *Future enhancements*

In order to provide a more complete view of the overall PRACE systems status, the general view provided to all PRACE users should include:

- A matrix of site availability through the network (i.e., provide an answer to the question: “is Site A reachable from Site B?”)
- A status report of the INCA monitoring system itself.

INCA can use dependency information on some tests, and –for instance– grey out all the entries of an unavailable site. Dependency upon the network availability and systems maintenance status will be targeted for the PRACE production operations phase.

# 7 Conclusions and future work

The work WP4 has done so far, has laid the grounds for the PRACE distributed infrastructure. Additional servers providing central services necessary for the operation of the infrastructure are in place for Grid (UNICORE 6) and interactive (GSI-SSH and X.509-SSH) access, for data movement within the ecosystem (through Globus RFT/GridFTP), and for user administration and report on resource usage. A monitoring infrastructure has been defined and sample reports are already in place, together with a complete role-based access model. Configuration specifications are provided to WP5 to enable the services on the prototypes, together with references and pointers to further information and details.

Some issues emerged during the discussions and preparatory work on the enhanced software stack: differing security policies can lead to long negotiations and delays in deciding and deploying a solution. It seems likely that, in order to come to a common specification, trust must be built among the participating centres; emergency procedures must be agreed upon; and a common understanding about the concessions that local security policies have to make must be reached. It is also quite likely that feedback from Industry users will have an impact on the assessment and provision of common security measures. WP4 expects to tackle this issue by promoting (together with WP2 and other interested parties) a security forum where experts and decision makers from the different sites meet and agree on a joint security policy.

While each piece of software has been evaluated to make the selection described in this deliverable, further evaluation of the *ensemble* is needed, in order to assess how the whole stack would perform under PRACE production conditions. WP4 has started preparation work for this evaluation; services will be tried out from a „user“ perspective when they will be available on the different PRACE prototypes. Within this context, a refinement of the proposed monitoring infrastructure is expected, based on actual usage, operational model, close-to-production assessment, and feedback from the operations teams at sites.

Based upon the close-to-production evaluation, the operational model finally chosen by PRACE, input from the interested Industry users and feedback from other PRACE WPs, the software stack will be adapted and further enhanced. The final software stack installed on the prototypes will be documented in D4.1.4. A separate specification will be produced in D4.3, fully describing the set-up of the PRACE production infrastructure, and making recommendations for further developments, especially taking into account new developments, from DEISA.