SEVENTH FRAMEWORK PROGRAMME
Research Infrastructures

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

CAPACITIES

PRACE-3IP

PRACE Third Implementation Phase Project

Grant Agreement Number: RI-312763

D6.3.1
First Annual Technology Report

Final

Version: 1.0
Author(s): Michael Rambadt, FZJ
Date: 23.8.2013

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: < **D6.3.1**>	
	Deliverable Nature: <DOC_TYPE: Report>	
	Deliverable Level: PU *	**Contractual Date of Delivery:** 31 / August / 2013
		Actual Date of Delivery: 31 / August / 2013
	EC Project Officer: Leonardo Flores Añover	

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

Document Control Sheet

Document	**Title: First Annual Technology Report**	
	ID: D6.3.1	
	Version: <1.0 >	**Status:** ***Final***
	Available at: http://www.prace-project.eu	
	Software Tool: Microsoft Word 2007	
	File(s): D6.3.1.docx	
Authorship	**Written by:**	Michael Rambadt, FZJ
	Contributors:	Gabriele Carteni, BSC Ilya Saverchenko, LRZ Zoltan Kiss, NIIF Frank Scheiner, HLRS Ralph Niederberger, FZJ Mateo Lanati, LRZ Andrew Turner, EPCC Jules Wolfrat, SURFsara
	Reviewed by:	Andreas Schott, RZG; Dietmar Erwin, FZJ
	Approved by:	MB/TB

Document Status Sheet

Version	**Date**	**Status**	**Comments**
0.1	08/July/2013	Draft	TOC
0.2	05/August/2013	Draft	Contributions for Bolt, gsatellite, GridFTP repository
0.3	12/August/2013	Draft for internal review	Contributions for gsatellite (update),

			perfSONAR, DMOS
1.0	23/August/2013	Final	Update for Bolt, gsatellite

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.

Copyright notices

Table of Contents

References and Applicable Documents

[1] BOLT, https://github.com/aturner-epcc/bolt

[2] UNICORE, Uniform Interface to Computing Resources, http://www.unicore.eu/

[3] PRACE-1IP D6.2 “First annual report on the technical operation and evolution”, http://prace-ri.eu/IMG/pdf/D6-2_1ip.pdf

[4] https://github.com/fr4nk5ch31n3r/gsatellite/wiki

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

[6] http://git-scm.com/

[7] http://www.deisa.eu

[8] http://www.prace-project.eu/IMG/pdf/d7.4_3ip.pdf

[9] http://geant3.archive.geant.net/Services/NetworkPerformanceServices/Pages/perfSONARMDM.aspx

[10] http://datatracker.ietf.org/wg/ipsec/charter/

[11] http://www.prace-project.eu/IMG/pdf/D6-1_2ip.pdf

[12] http://www.geant.net/About/Pages/default.aspx

[13] http://www.geant.net/Services/ConnectivityServices/Pages/GEANTPlus.aspx

List of Acronyms and Abbreviations

AAA	Authorization, Authentication, Accounting
AMD	Advanced Micro Devices
BSC	Barcelona Supercomputing Center (Spain)
CINECA	Consorzio Interuniversitario, the largest Italian computing centre (Italy)
DEISA	Distributed European Infrastructure for Supercomputing Applications. EU project by leading national HPC centres.
DMOS	Distributed Maintenance Information Organisation System
EPCC	Edinburg Parallel Computing Centre (represented in PRACE by EPSRC, United Kingdom)
FTP	File Transfer Protocol
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
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	Certificate based File Transfer Protocol
HPC	High Performance Computing
HLRS	Höchstleistungsrechenzentrum Stuttgart (Germany)
ICHEC	Irish Centre for High-End Computing
IDRIS	Institut du Développement et des Ressources en Informatique Scientifique (represented in PRACE by GENCI, France)
Inca	Service monitoring tool used in PRACE
IPsec	Internet Protocol Security
iRODS	integrated Rule-Oriented Data-management System, a community-driven, open source, data grid software solution
ISTP	Internal Specific Targeted Project
LRZ	Leibniz Supercomputing Centre (Garching, Germany)
MPI	Message Passing Interface
NIIF	Nemzeti Információs Infrastruktúra Fejlesztési Intézet (National Information Infrastructure Development Institute, Hungary)
OpenMP	Open Multi-Processing
Tier-0/-1	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
UiO	University in Oslo
UNICORE	Uniform Interface to Computing Resources. Grid software for seamless access to distributed resources.
XSEDE	Extreme Science and Engeneering Discovering Environment (NSF-funded US-Project)

Executive Summary

Task 6.3 ‘Technical evolution of the PRACE operational services’ in work package 6 ‘Operation of the Distributed Research Infrastructure’ of PRACE-3IP continues the work from PRACE-1IP (WP6, Task 6.3). It also takes over the technical evolution of the PRACE operational services from work package 10 ‘Advancing the Operational Infrastructure’ in PRACE-2IP after the completion of WP10.

In the first year 6.3 started to evaluate the *Bolt* software as a job submission tool complementing UNICORE.

Also the evaluation of *gsatellite* has been started to provide additional functionality to GridFTP, the standard file transfer technology in PRACE.

For PRACE internal needs the task members worked on and finished the setup of an internal *GridFTP repository*. Also the evaluation of *DMOS* is almost finished. Additionally the *perfSONAR* evaluation has been started.

A further objective was the handover from WP10. This has been planned and agreed in detail with the WP10 members.

1 Introduction

The objectives of WP6.3 in the first year were:

- Technology watch for the complementation of existing PRACE services
- Working on PRACE internal software needs

In the first year the team members concentrated on technology watch for the complementation of existing PRACE services and on PRACE internal needs coming from other work packages.

Concerning the technology watch, experiences coming from the local user support teams at each site were taken into account.

This is the case for *Bolt* and *gsatellite* (see chapter 2 and 3). Both tools were known and used already at respectively one PRACE site, and were considered to complement the PRACE core services UNICORE and GridFTP.

The *GridFTP repository* is one of the PRACE internal (WP7) needs and is described in chapter 4. Topics of chapter 5, and 6 cover the *DMOS* and *perfSONAR* evaluation, respectively.

Chapter 7 outlines the planned work for the second year.

2 Bolt evaluation

Bolt [1] is a platform-independent tool for generating job submission scripts that presents a common command line based interface to different batch systems to the user. Thereby it can optimize the use of the underlying architecture of the target system. The tool, which is developed and maintained by EPCC, can also generate cross-platform scripts for architectures that are different from the local resource.

The tool will be an addition to the “Uniform access to HPC” service, that is currently supported by the UNICORE [2] software.

UNICORE already provides a common and uniform interface to distributed computing resources in a more sophisticated and featured way including job management, data staging and resource discovering.

Bolt on the other hand is a more lightweight solution that addresses users who prefer to interact with a batch system directly without the overhead of UNICORE.

2.1 Reasons, benefits and constraints

Each HPC resource in PRACE (Tier-0 and Tier-1) has a unique setup for batch job scheduling, for assigning parallel tasks, and shared-memory threads to hardware resources. Even if some sites are adopting the same scheduler and resource manager the local configurations and limits differ substantially. This variation makes it difficult for end users to quickly generate correct job submission scripts.

Another issue is the increasing complexity of the underlying processor hardware. That requires a detailed understanding of the processor architecture to decide on the optimal pinning arrangement for their job.

For example, the current AMD Bulldozer architecture (used by a number of Cray sites in PRACE) shares a floating-point unit between two cores; performance improvements are available to many codes by half-populating nodes and pinning tasks to every other core. This allows the code to exploit a double-width floating-point unit and increase both memory and interconnect bandwidth per task. Specifying the correct options to the parallel job launcher command for an optimal distribution of the parallel tasks is non-trivial.

With Bolt the underlying processor technology can be encoded in a HPC resource description configuration file to produce the most efficient distribution of tasks and threads within the constraints of the user’s job specification.

The benefits of Bolt are:

- Common command line interface across PRACE execution systems for generating job submission scripts
- Efficient using of distributed resources
- Increasing code performance due to more optimal task/thread placement
- Sites do not have to change their scheduling/resource management procedures
- Minimal installation effort.

2.2 Evaluation status

CINECA, EPCC, ICHEC, and LRZ work on the Bolt evaluation.

The evaluation started with the definition of a test bench. The test bench includes two Tier-0 systems and two Tier-1 systems. These have three different batch systems (Torque/Moab, LoadLeveler and PBSPro):

- CINECA (Italy): Tier-0 system “FERMI” (LoadLeveler)
- EPCC (UK): Tier-1 system “HECToR” (PBSPro)
- ICHEC (Ireland): Tier-1 system “Stokes” (Moab/Torque)
- LRZ (Germany): Tier-0 system “SuperMUC” (LoadLeveler)

In addition to the test bench also the tests that are expected to run at all involved sites were defined:

- Submission of a MPI parallel job using a single node

- Submission of a MPI parallel job using more than one node
- Simple OpenMP job using all 2 threads
- Simple MPI parallel job using more tasks than available on the system
- Simple MPI parallel job with a walltime that exceeds the maximum available walltime on the system
- OpenMP job using more threads than available on a node

The Bolt tool has been adapted to the test bench systems with the respective configuration files. First tests by EPCC have been successful. Currently, the local test site representatives are executing the defined tests.

During the evaluation several feature requests have been generated to extend the Bolt functionality. The requests have been submitted to the Bolt developers.

2.3 Planning of the second year

The ISTP document [3] is a PRACE internal guide for evaluating new technologies and is described in detail in the PRACE-1IP WP6 D6.2 “First annual report on the technical operation and evolution”. The document has been changed during the *Bolt* evaluation phase. The ISTP now requires a security audit by the PRACE security forum at the beginning of each service evaluation. Therefore 6.3 have forwarded the *Bolt* security check request to the security forum. Results are expected for September 2013.

If the security forum has no objections, the evaluation of Bolt will be continued and extended. It is planned to include further batch systems, like SLURM and LSF.

3 Gsatellite evaluation

gsatellite [4] is an open source data transfer scheduling and managing tool. On top of the *gtransfer* [5] tool (see PRACE-2IP D6.3 “Second Annual Operations Report of the Tier-1 Service”), it allows users to submit and manage large GridFTP data transfers running non-interactively in the background.

The evaluation process consists of a planning phase, the setup of test environments including the relevant network connections, and the detailed testing of *gsatellite*, especially concerning the needs of the PRACE community.

The partners in this task are HLRS and NIIF.

3.1 Reasons, benefits and constraints

Currently, PRACE users using GridFTP need to perform data transfers, including error handling, manually and to monitor them from the beginning to the end. This could get a rather expensive task, especially in case of:

- transfers of big data that will last a long time
- regularly scheduled transfers (e.g. every Friday at 3 pm)
- transfers that have to be done within a specific timeframe (e.g. only after 10 pm and before 6 am)

gsatellite provides this functionality. Additionally, *gsatellite* makes the transfers reliable and thus retries them automatically in case of any error. Users can log into one or multiple

(frontend) machines, submit their data transfer jobs, leave, and let *gsatellite* take care of the remaining work.
They do not have to be online during the data transfer but can return at any time and check the status of their jobs. If required also the email notification can be activated to get information about the transfer status.

3.2 Evaluation status

The evaluation started in June 2013 and is done in three phases

- Planning phase
- Setup of the test environment
- Detailed testing of *gsatellite*

Within the initial planning phase the requirements for the test environment regarding needed additional software and network connectivity were defined.

The requirements for the test environment are as follows:

- The system must be able to send status emails
- The *git* [6] versioning system has to be supported
- A login for all team members must be possible on all test machines
- The team members can write to a shared directory and execute predefined gsatellite services, e.g. email notifications
- Each test machine has a connection to both the PRACE internal network and to the public internet
- Each test machine has a connection to the PRACE GridFTP servers

The test environment at NIIF has been set up in the meantime still expecting connection to the PRACE internal network. As soon as the PRACE internal network will be available, the test machines will be connected as well, and NIIF will start with the *gsatellite* tests.

3.3 Planning of the second year

In year 2 it is planned to add three additional sites and to execute the following detailed *gsatellite* tests/evaluations:

- Installation process and compatibility with different operating systems and software environments
- General usage and user experience
- General stability and resource requirements
- Data transfer benchmark tests
- Detailed testing of the gsatellite features

4 DMOS evaluation

Distributed Maintenance Information Organisation System, *DMOS*, is a tool to announce and manage service and resource maintenance information. The tool provides functionality for management, for instance for documentation and distribution, and persistent storage of information describing planned downtime of resources and services in a federated distributed environment.

The *DMOS* development started during DEISA2 [7] to provide an easy to use and flexible solution for management of maintenance information. Thus it should replace the DEISA Wiki that was used to store information about temporary unavailable DEISA resources.

4.1 Reasons, benefits and constraints

The PRACE operations team relies on a dedicated section in the PRACE Wiki for announcement and documentation of service and resource maintenances. Each partner is instructed to publish information about scheduled and unplanned maintenances in this Wiki section. Maintenances can be announced at a site, resource or, in special cases, service level and contain the following details:

- Scheduled start of a maintenance
- Scheduled end of a maintenance
- Time of the actual maintenance ending
- Affected sites and resources
- Maintenance description

The provided information describes the general availability of services and resources and allows PRACE operations team to plan support, deployment and maintenance activities. Furthermore this information helps PRACE Operator on Duty to appropriately react to problems with the e-Infrastructure, for instance by appropriately treating service failures caused by unavailability of the respective resource.

Maintenance information collected in PRACE Wiki has been turned out as not sufficient for a detailed analysis of the real time e-Infrastructure state. For example:

- Creation and update of maintenance announcements can only be performed manually
- Information provided in the maintenance description field is often not suited for end-users
- Service and resource dependencies cannot be specified in the PRACE Wiki

DMOS addresses these limitations and provides added flexibility by supporting standard access interfaces and extended functionality. *DMOS* implements a multi-tier architecture and comprises two main components: a backend and a graphical user interface. The backend includes a relational database and implements the necessary logic and interfaces for data access and processing. The graphical user interface is provided by a web application that supports authentication, authorization and a variety of data views that focus on interests and requirements of PRACE stakeholders.

4.2 Evaluation status

DMOS has been developed by the two PRACE partners IDRIS (France) and LRZ (Germany). Both sites also were involved in running the *DMOS* testbed and have been supported by UiO (Norway).

The *DMOS* functionality was successfully evaluated against the following operational requirements:

- Provide information about scheduled maintenances
- Store information in a format suitable for manual and automatic processing
- Contain information about sites, resources and services
- Support several message granularity levels
- Implement notification functionality

- Offer persistent data storage
- Expose interfaces to other tools and services, for instance monitoring, user support, etc.

The evaluation results were presented and discussed with PRACE operations team and *DMOS* was passed on for deployment in a production environment.

4.3 Planning of the second year

Currently the evaluation and planning of *DMOS* integration with other production tools and services used for PRACE operations, such as the Inca monitoring, is ongoing.

5 GridFTP repository for WP7

Task 6.3 worked together with WP7 to provide a long term file repository based on GridFTP to store the PRACE Unified Benchmarks Suite [8] including, programs, input data and results.

The requirements included the possibility to access the storage server from user workstations, so from the public Internet and full connectivity to Tier-0 and Tier-1 systems on the dedicated PRACE network. The envisaged dataset was made up of few large files, between 100 and 200 GB, plus a limited number of smaller files.

5.1 Evaluation status

LRZ set up a server (gridmuc.lrz.de) with two network interfaces, one to the general Internet and one to the PRACE network, and with up to 1 TB of disk space for file hosting.

GridFTP access has been granted to all PRACE staff members listed in the PRACE LDAP server. The tests of this service were successfully completed.

5.2 Planning of the second year

The activity has been finished and the *GridFTP repository* is ready to use.

6 perfSONAR evaluation

Network management tools were set up already in the beginning of PRACE to monitor the dedicated network infrastructure. In this context PRACE also evaluated the former version of the GÉANT *perfSONAR* MDM services [9]. But it turned out that this *perfSONAR* version did not have significant advantages for PRACE. It was agreed to test these services again as soon as a new release with additional features will be available.

End of 2012 GÉANT announced this new release. In December 2012 PRACE representatives discussed the possible future integration into the PRACE network monitoring with the *perfSONAR* product manager. In March 2013 PRACE experts participated in a *perfSONAR* workshop.

The contributor to this task is FZJ. In the meantime a first *perfSONAR* server system has been set up at FZJ and the tests have been started.

The tests will be continued in year 2. A detailed report will follow in the next deliverable.

7 Planning for the second year

In addition to the on-going activities that have been reported in the previous chapters, three evaluation topics have been added for the second year. On the one hand this is the continuation of the WP10 work of PRACE-2IP and the evaluation of the *GÉANT Plus service*. On the other hand it is about the planned collaboration between the European and the US supercomputing initiative, PRACE and XSEDE.

7.1 Continuation of the work from WP10 of PRACE-2IP

WP10 evaluated potential new operational services in PRACE-2IP. WP6.3 will continue the work started and proposed for deployment by WP10:

- Improving the existing infrastructure, concerning AAA-GridSafe, Storage Accounting and reporting, DECI-Portal, DECI-Project management-DB, PRACE-Information Portal, GOCDB, Inca improvements, and the collaboration with other technologically oriented projects
- Data-Services with the continuation of the new file transfer technology, the iRODS, and the file system technologies evaluation

Details can be found in the final deliverable D10.2 “Second Annual Report of WP10 – Progress on Technology Scouting and Development” of WP10.

7.2 Evaluation of the GÈANT Plus service

PRACE-2IP implemented the *IPsec* [10] solution to provide newTier-1 sites with an alternative, less cost-intensive connection to the PRACE internal network. Details to *IPsec* can be found in the PRACE-2IP WP6 deliverables D6.1 “First Annual Operations Report of the Tier-1 Service” [11] and D6.3 “Second Annual Operations Report of the Tier-1 Service”.

During the IPsec implementation GÉANT [12], the pan-European research and education network, has announced the *GÉANT Plus service* [13]. It allows user access to point-to-point circuits of between 100Mbit/s and 10Gbps across an existing pre-provisioned network.

Since this approach is similar to the IPsec one, it has been decided to start a subtask in 6.3 to test these services, too. NIIF (Hungary) offered to connect their HPC environment via the *GÉANT Plus service* instead of the IPsec solution. Therefore GÉANT has been contacted already to provide a *GÉANT Plus* connection between Budapest and Frankfurt. It is assumed that this link will be available end of 2013.

7.3 PRACE XSEDE provision of collaborative access

XSEDE, the US supercomputing initiative, and PRACE have a number of common challenges. Hence, PRACE and XSEDE representatives decided to start focused collaborations.

The respective contact between the PRACE and the XSEDE operation team members has been established already, and in a first phone conference the following topics have been defined for future common work:

- Support for integrated environments and services
- Offer to preparatory access on PRACE and XSEDE resources (the exact conditions and available resources still have to be defined)
- Involve existing peer review mechanism

As one result a joint call for PRACE/XSEDE user communities is planned to provide them with common computing time.