SEVENTH FRAMEWORK PROGRAMME
Research Infrastructures

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

PRACE-3IP

PRACE Third Implementation Phase Project

Grant Agreement Number: RI-312763

D6.2.2
Second Annual Integration Report

Final

Version: 1.0
Author(s): Javier Bartolome, BSC
Date: 17.6.2014

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: < D6.2.2>	
	Deliverable Nature: <DOC_TYPE: Report / Other>	
	Deliverable Level: PU *	Contractual Date of Delivery: 30 / June / 2014
		Actual Date of Delivery: 30 / June / 2014
	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: Second Annual Integration Report	
	ID: D6.2.2	
	Version: 1.0	Status: *Final*
	Available at: http://www.prace-project.eu	
	Software Tool: Microsoft Word 2007	
	File(s): D6.2.2.docx	
Authorship	Written by:	Javier Bartolome, BSC
	Contributors:	Ralph Niederberger, JSC Philippe Collinet, IDRIS Stelios Erotokritou, CaSToRC Zoltan Kiss, NIIF Josip Jakic, IPB Niall Wilson, ICHEC Filip Stanek, VSB-TUO
	Reviewed by:	Mark Bull, EPCC Thomas Eickermann, JSC
	Approved by:	MB/TB

Document Status Sheet

Version	Date	Status	Comments
0.1	14/April/2014	Draft	TOC
0.2	17/April/2014	Draft	Introduction and chapter 2 edited.
0.3	23/April/2014	Draft	Network services included
0.4	30/April/2014	Draft	VSB-TUO contribution

			added
0.5	30/April/2014	Draft	IDRIS contribution added
0.6	07/May/2014	Draft	CastorC contrib added
0.7	14/May/2014	Draft	IPB/ICHEC contrib added
0.9	10/June/2014	Draft for internal review	
0.99	16/June/2014	Final Draft	Included review comments
1.0	17/June/2014	Final version	

Document Keywords

Keywords:	PRACE, HPC, Research Infrastructure

Disclaimer

This deliverable has been prepared by the responsible Work Package of the Project in accordance with the Consortium Agreement and the Grant Agreement n° RI-312763. It solely reflects the opinion of the parties to such agreements on a collective basis in the context of the Project and to the extent foreseen in such agreements. Please note that even though all participants to the Project are members of PRACE AISBL, this deliverable has not been approved by the Council of PRACE AISBL and therefore does not emanate from it nor should it be considered to reflect PRACE AISBL's individual opinion.

Table of Contents

List of Figures

References and Applicable Documents

[1] http://www.prace-project.eu
[2] PRACE-3IP Deliverable D6.2.1, First Annual Integration Report. http://prace-ri.eu/IMG/pdf/d6.2.1.pdf

List of Acronyms and Abbreviations

AAA	Authorization, Authentication, Accounting.
AISBL	Association International Sans But Lucratif (legal form of the PRACE-RI)
AMD	Advanced Micro Devices
ATI	Array Technologies Incorporated (AMD)
BSC	Barcelona Supercomputing Center (Spain)
CGI	Common Gateway Interface
CPU	Central Processing Unit
CUDA	Compute Unified Device Architecture (NVIDIA)
DDN	DataDirect Networks
DDR	Double Data Rate
DMA	Direct Memory Access
DP	Double Precision, usually 64-bit floating point numbers
EC	European Commission
EP	Efficient Performance, e.g., Nehalem-EP (Intel)
EPCC	Edinburg Parallel Computing Centre (represented in PRACE by EPSRC, United Kingdom)
EPSRC	The Engineering and Physical Sciences Research Council (United Kingdom)
EX	Expandable, e.g., Nehalem-EX (Intel)
FC	Fiber Channel
FP	Floating-Point
FPGA	Field Programmable Gate Array
FPU	Floating-Point Unit
FZJ	Forschungszentrum Jülich (Germany)
GB	Giga (= 2^{30} ~ 10^9) Bytes (= 8 bits), also GByte
Gb/s	Giga (= 10^9) bits per second, also Gbit/s
GB/s	Giga (= 10^9) Bytes (= 8 bits) per second, also GByte/s
GCS	Gauss Centre for Supercomputing (Germany)
GDDR	Graphic Double Data Rate memory
GÉANT	Collaboration between National Research and Education Networks to build a multi-gigabit pan-European network, managed by DANTE. GÉANT2 is the follow-up as of 2004.
GENCI	Grand Equipement National de Calcul Intensif (France)
GFlop/s	Giga (= 10^9) Floating point operations (usually in 64-bit, i.e. DP) per second, also GF/s
GHz	Giga (= 10^9) Hertz, frequency =10^9 periods or clock cycles per second
GigE	Gigabit Ethernet, also GbE
GNU	GNU's not Unix, a free OS
GPGPU	General Purpose GPU
GPU	Graphic Processing Unit
HP	Hewlett-Packard
HPC	High Performance Computing; Computing at a high performance level at any given time; often used synonym with Supercomputing
IB	InfiniBand
IBM	Formerly known as International Business Machines
ICE	(SGI)

IDRIS	Institut du Développement et des Ressources en Informatique Scientifique (represented in PRACE by GENCI, France)
I/O	Input/Output
JSC	Jülich Supercomputing Centre (FZJ, Germany)
KB	Kilo (= 2^{10} ~10^3) Bytes (= 8 bits), also KByte
MB	Mega (= 2^{20} ~ 10^6) Bytes (= 8 bits), also MByte
MB/s	Mega (= 10^6) Bytes (= 8 bits) per second, also MByte/s
MFlop/s	Mega (= 10^6) Floating point operations (usually in 64-bit, i.e. DP) per second, also MF/s
MHz	Mega (= 10^6) Hertz, frequency =10^6 periods or clock cycles per second
Mop/s	Mega (= 10^6) operations per second (usually integer or logic operations)
Open MP	Open Multi-Processing
OS	Operating System
PCPE	PRACE Common production environment
PRACE	Partnership for Advanced Computing in Europe; Project Acronym
QDR	Quad Data Rate
RAM	Random Access Memory
RDMA	Remote Data Memory Access
RISC	Reduce Instruction Set Computer
RPM	Revolution per Minute
SGI	Silicon Graphics, Inc.
SMP	Symmetric MultiProcessing
SP	Single Precision, usually 32-bit floating point numbers
TB	Tera (= 240 ~ 1012) Bytes (= 8 bits), also TByte
TFlop/s	Tera (= 1012) Floating-point operations (usually in 64-bit, i.e. DP) per second, also TF/s
Tier-0	Denotes the apex of a conceptual pyramid of HPC systems. In this context the Supercomputing Research Infrastructure would host the Tier-0 systems; national or topical HPC centres would constitute Tier-1
UNICORE	Uniform Interface to Computing Resources. Grid software for seamless access to distributed resources.
UV	Ultra Violet (SGI)

Executive Summary

This report describes the activities of Task 6.2 (WP6): the integration and upgrade of new Tier-0 and Tier-1 systems during the second year of PRACE 3-IP.

A comprehensive procedure for monitoring and reporting the upgrade of a system in PRACE has been designed and tested. This procedure for the upgrade of a system must guarantee, as far as possible, an undisturbed migration of the PRACE services, with minimal impact on the services at other sites. It has been designed to be applied both for Tier-0 and Tier-1 systems. The procedure was successfully established during the first year, and improved during this second year for upgrading Turing at IDRIS, Fionn at ICHEC and ARCHER at EPCC.

The same procedure, with minor changes, has been used for monitoring and reporting the integration of new Tier-1 sites, which has helped to standardize these changes in the infrastructure. During this second year integration of NIIF, CaSToRC, VSB-TUO has been completed successfully, and the integration of IPB has started and is expected to be completed soon.

The experience from these upgrades and integrations has been used for improving the procedure itself. The upgrade/integration procedure, as discussed in this document, consists of nine steps that should cover all possible scenarios.

1 Introduction

The existing distributed Tier-0 and Tier-1 infrastructure is regularly extended with the addition of new Tier-0 and Tier-1 resources, as well as by upgrades of existing systems. This task has been responsible for the integration and upgrade of new Tier-0 and Tier-1 sites, for which a procedure has been defined. The description of the procedure and the results of using it are all documented in Section 2 of this document.

In Section 3 the integration of new Tier-0/Tier-1 systems is described. During this second year, systems from NIIF in Hungary, CaSToRC in Cyprus, VSB-TUO in Czech Republic and IPB in Serbia have been integrated as Tier-1 sites.

In Section 4 the upgrade of existing Tier-0/Tier-1 systems is described. In this second year Turing from IDRIS (France), Fionn from ICHEC (Ireland) and Archer from EPCC (UK) have been upgraded and integrated as Tier-1 machines.

This document is of interest for system administrators and other staff involved in the management of systems. It gives details on how the installation or upgrade of a system in the PRACE infrastructure can be managed in a controlled way.

2 Upgrade procedure for Tier-0 and Tier-1 systems

The Change Management procedure, available since PRACE-1IP, is used to manage all changes that have an impact on the PRACE infrastructure. This does not give a good overview of all the changes involved with a system replacement and especially not of all the dependencies. This is not surprising because it was conceived for managing changes on services, not on systems. Since there will also be many similarities between different system replacements, a general guideline for the replacement of a system was needed.

This procedure, defined during the first year of PRACE 3-IP, has been improved upon during this second year, and its use has been generalized in the integration and upgrade of all new Tier-0/Tier-1 systems.

The procedure was designed to ensure continuous service to users and assumes that:

- The system being decommissioned has been withdrawn on time from any PRACE or DECI call and is free of project allocations or;
- Active project allocations, if any, are migrated to other systems (or to the new system) with a minimum of downtime;
- Details about the procurement process of the new system are not relevant for the operational integration.

The most important requirement is that PRACE services be available on the new system at the time that project allocations should be active on the new system.

The “System Upgrade Procedure” is based on 9 phases (which can overlap):

1. Plan and Prepare the System Upgrade (a schedule is produced)
2. Internal Announcements
3. Site Preparation
4. System Installation and Acceptance Testing
5. Installation of PRACE services on the new system
6. Pre-production
7. Shutdown of PRACE Services on the old system (some services might not be affected) and start of production on the new system
8. Old System Shutdown

9. Review and Process closing

More detailed descriptions of each of the phases of the upgrade procedure are described in the first annual report [2], and most of the different phase descriptions are still valid today. The improvements and changes to the procedure have been made in the template document for upgrades in the internal PRACE wiki.

3 Integration of new Tier-0 and Tier-1 systems

The procedure for integrating new Tier-0 and Tier-1 sites is the same as that of a system upgrade, except that steps dealing with "old system" do not apply. In this project period only new Tier-1 systems have been integrated. These are discussed in the following sections.

3.1 NIIF

The NIIF supercomputer is a fat-node HP cluster which is a very sophisticated type of blade technology (CP4000SL). It uses the latest AMD Opteron 6174 type processors with 12-core Magny-Cours. The total number of cores is 768. The interconnect network is a redundant QDR Infiniband network. The whole system has 2 TByte memory and the computing power is 5.4 TFlop/s. A water-cooled rack was placed in the system to increase the energy efficiency. This unique supercomputer runs very effectively using mixed-mode parallel programming paradigms, and each node is a powerful 24 cores SMP computer.

Status of PRACE Services installation

NIIF has been successfully connected to the PRACE network via Geant Plus VPN Layer 2 services and network monitoring has been setup successfully. The following actions were taken to complete the integration:

- Established PRACE connection using GEANT+ services, managing outages
- Configured and installed network monitoring tools
- Opened middleware access for PRACE sites
- Prace_service script installed, modified, change management created, managed
- UNICORE migration to SLURM scheduler completed, configured with PRACE-LDAP
- PRACE accounting CGI installed
- Authorization-management completed using own PRACE-LDAP scripts
- Inca configuration completed using ssh method
- Completed PCPE environment creation
- Site specific documentation created, migration and further translation of the services provided for PRACE on a standard, wiki-based infrastructure, updated docs links at PRACE SVN

3.2 CaSToRC

Figure 1 - Cy-Tera Supercomputer

Cy-Tera - the computational resource of CaSToRC, was installed at its premises in November 2011 with its acceptance tests finalised in April 2012. Cy-Tera is a IBM iDataplex cluster of 35TFlops with 1392 cores of Intel Westmere processors and GPUs based on NVIDIA M2070.

Since May 2012 it has been accepting production level users from LinkSCEEM/Cy-Tera production calls. The operations and suppport group of CaSToRC is relatively new, but through these production calls has gained the sufficient experience to handle the continuous operation of Cy-Tera and also acquired the capability to promptly handle user requests for software project requirements and other user issues too. Because of the above, CaSToRC was well prepared to handle the integration of Cy-Tera to DECI for 10% of its available annual resources.

Status of PRACE Services installation

For the integration to occur, CaSToRC had to join PRACE's dedicated network infrastructure. Since The Cyprus Institute (where CaSToRC is based) is not connected to GEANT, this integration took place using IPsec. Through the cooperation of CaSToRC and PRACE NOC staff, a small IP address range was reserved for the sole purposes of PRACE usage, and the IPSEC router was set up to connect CaSToRC to the Julich PRACE site via the PRACE switch in Frankfurt, and later configured to allow for connectivity between CaSToRC and other PRACE partners. Staff from other PRACE partners were also contacted, informing them of CaSToRC's intergration to the PRACE dedicated network, so they could in turn configure their network to allow for connectivity between CaSToRC and their respective sites, and network monitoring was finally setup successfully.

Once CaSToRC network integration took place, installation of PRACE services upon Cy-Tera began based on four main categories, namely Data Services, Compute Services, AAA services and User Services.

Implementation of Data Services included the setup of a GridFTP server on CyTera's login node, to enable high-performance, secure and reliable data transfers, as well as the installation of the Gtransfer tool, which was developed internally by PRACE.

Compute Services dealt with the integration of UNICORE with the local batch system, which is Torque/Moab for Cy-Tera. All components of UNICORE, i.e. Gateway, Unicore/X,

XUUDB and TSI were installed and configured in order to allow for job submission to Cy-Tera through UNICORE.

AAA services were also setup with the creation of the CaSToRC LDAP branch, in conjunction with the implementation of LDAP related tools, to enable the addition of new users on CaSToRC's branch on the PRACE LDAP server. GSI-SSH has also been installed and made available to enable interactive user access to Cy-Tera. Furthermore, the accounting CGI script has been set up with a MySQL database.

User Services included the implementation of the common PRACE modules enviroment and prace_service script. CaSToRC specific user documentation is also under preparation.

Finally, in order to monitor the PRACE network and services, CaSToRC has been integrated into Inca and PRACE network monitoring mechanisms, thus all provided services and tools are constantly monitored.

It should be noted that Cy-Tera was part of DECI 11 with one project from Poland allocated to CaSToRC. Since March 2014 user accounts have been created and project jobs are running upon Cy-Tera without any user issues.

3.3 VSB-TUO

The Anselm cluster of VSB-TUO consists of 209 compute nodes, totalling 3,344 compute cores with 15 TB of RAM and giving over 94 TFlop/s theoretical peak performance. Each node is a powerful x86-64 computer, equipped with 16 cores, at least 64 GB of RAM, and a 500 GB hard drive. All nodes are interconnected by a non-blocking fat-tree QDR Infiniband network and equipped with Intel Sandy Bridge processors. A few nodes are also equipped with NVIDIA Kepler K20 GPU (23 nodes) or Intel Xeon Phi 5110p MIC accelerators (4 nodes).

The cluster runs bullx Linux operating system, which is compatible with the RedHat Linux family. A user data shared file-system (HOME, 320 TB) and a job data shared file-system (SCRATCH, 146 TB) are available to users. The PBS Professional workload manager provides computing resources allocations and job execution.

Status of PRACE Services installation

VSB-TUO has been successfully connected to the PRACE network via an IPSEC-GRE tunnel solution. Network monitoring has been setup successfully.

Because the system was installed and put into production as "new Tier-1" system in PRACE 2IP, in PRACE-3IP only upgrade or installation of additional services was done as part of PRACE integration.

During this second year, we can highlight the installation of the following PRACE services:

- User documentation completed.
- Inca monitoring installed.
- GridFTP installed.
- Gtransfer installed.
- Prace_service script and config installed.

3.4 IPB

The PARADOX cluster at the Scientific Computing Laboratory of Institute of Physics Belgrade consists of 106 HP Proliant SL250s compute nodes, with 1,696 cores and 3.3 TB of

RAM in total, with 106 TFlop/s theoretical peak performance. Each node consists of two 8-core Sandy Bridge Intel Xeon E5-2670 processors with 32 GB of RAM, one NVIDIA Tesla M2090 accelerator card with 5,375 MB of RAM, 512 CUDA cores at 1.3GHz with Compute capability 2.0 and a 500 GB local hard disk. Nodes are interconnected by a QDR InfiniBand network.

The operating system on PARADOX cluster is Scientific Linux 6.4. A Lustre shared file system (/home) and a local file system on each worker node (/scratch), are available to users. The batch system managing resource allocation, job submission and execution is Torque with Maui scheduler.

Status of PRACE Services installation

The following actions were taken as part of the integration:

- Site representatives and System information were added to the general list.
- Staff access to the internal services (Wiki, TTS, BSCW, SVN) was verified and general operational procedures were acknowledged.
- Site branch in the LDAP server was applied for and was created.
- PCPE environment has been installed.
- PARADOX Cluster User Guide initial version was created (http://www.scl.rs/PARADOXClusterUserGuide/).

Planned activities for the integration:

- Connection to the PRACE internal network is planned but its implementation time is still uncertain. Installation of the rest of production services, as defined by the Template for the integration of a new Tier-1 site, is planned to be finished by the end of July 2014.

4 Upgrades of Tier-0 and Tier-1 systems

There have only been upgrades of Tier-1 systems in this period.

4.1 Turing - IDRIS

The new IDRIS Tier-1 system Turing is a 4-rack IBM BlueGene/Q system. It replaced the 10-rack IBM BlueGene/P Babel system. Turing is connected to the 10Gbit/s Network and internet connectivity is provided through a login node. Although it is not scheduled to host DECI projects on this system, PRACE services have been installed nonetheless.

Status of PRACE Services installation

Network Services

- **PRACE Link:** Since network connectivity to IDRIS has been setup a long time ago, only local configuration changes had to be made. These have been implemented without any problems.

Data Services

- **GridFTP:** GridFTP has been installed and is fully operational since summer 2013. This service is also used by the PRACE staff in other on-going PRACE activities (iRODS task).
-

Compute Services

- **UNICORE:** UNICORE has been installed and has been operational since September 2013.

AAA Services

- **LDAP:** no change concerning this service, which was already available with the previous system.
- **GSISSH:** GSISSH has been installed and has been fully operational since summer 2013. This service is also used by the PRACE staff in other on-going PRACE activities (iRODS task).

User Services

- **Documentation:** user documentation for GridFTP, GSISSH and UNICORE is not up-to-date yet, but it will be ready in case of allocation of PRACE users.
- **PCPE:** the PCPE was integrated with the IDRIS module environment as soon as the system went into production.
- **RT-TTS:** queue for IDRIS is correctly available in the central PRACE TTS. User support is provided through the PRACE TTS and the mail prace-support@idris.fr.

Monitoring Services

- **Inca:** Inca was the last service to be installed since it is responsible for monitoring the status and health of all other PRACE services.

4.2 Fionn - ICHEC

The upgrade of the ICHEC Tier-1 system involved the replacement of the older Stokes system with a newly procured system call Fionn. Installation of this machine began in August 2013 and initial user access was granted in October 2013. DECI projects active on the Stokes system were migrated to Fionn in December prior to the decommissioning of Stokes. In summary, Fionn is a multi-architecture, hybrid system built to accommodate a broad range of requirements. The size and architecture of the constituent components is as follows:

- The Thin component, making up the bulk of the system, is an SGI ICE X system of 320 nodes with 7,680 cores (2.4 GHz Intel Ivy Bridge). Each node has 2x12-core processors, 64 GB of RAM and is connected using FDR InfiniBand. This amounts to 20 TB of RAM across the partition.
- The Hybrid partition contains 32 nodes, each with 2x10-core 2.2GHz Intel Ivy Bridge processors with 64 GB of RAM. This partition has accelerators from Intel and NVIDIA. 16 nodes have 32 Intel Xeon Phi 5110P while the other 16 have 32 NVIDIA K20X.
- The Fat section is an SGI UV2000 where 1.7 TB of RAM is accessible to 112 cores of Intel Sandy Bridge (14x8-cores processors) and 2 Intel Xeon Phi 5110P.
- The final component contains a set of service and administrative nodes to provide user login, batch scheduling, management, tape backup, switches, etc. Storage is provided via a DDN SFA12k-20 with 560TB of formatted capacity to all components of the machine via a Lustre filesystem.

Figure 2 - ICHEC Tier-1 Upgrade system - Fionn

Status of PRACE Services installation

The report on the planning and implementation of the upgrade for Fionn has been documented on the PRACE Operations wiki. The following is a summary of the main points.

Network Services

- **PRACE Link:** Fionn has not yet been connected to the PRACE network.

Data Services

- **GridFTP:** the GridFTP service was installed on Fionn in January 2014. As the system is not connected to the PRACE network, access to the service is firewall restricted to the designated PRACE door nodes.

Compute Services

- **UNICORE:** UNICORE has not been installed on Fionn as this is an on-request service and no DECI projects assigned to Fionn have requested this service.

AAA Services

- **LDAP:** LDAP branch for Fionn is operated by the central PRACE LDAP service and no changes were required.
- **GSISSH:** GSISSH access to Fionn was enabled in January 2014 and the prace-service configs updated. Access to the machine is also available via regular ssh with key exchange.

User Services

- **Documentation:** user documentation is up-to-date on the PRACE website. Documentation about specific PRACE services in the internal wiki is also up to date.
- **PCPE:** the Module environment has been installed following the internal PRACE specifications.
- **RT-TTS:** the queue for ICHEC is correctly available in the central PRACE TTS. User support is provided through the separate ICHEC TTS and the e-mail address prace-support@ichec.ie

Monitoring Services

- **Inca:** Inca monitoring was activated on Fionn in January 2014.

4.3 ARCHER - EPCC

ARCHER is an upgrade of the HECToR supercomputer in EPCC that finished production in March 2014. ARCHER is a Cray XC30 machine with the following hardware characteristics:

- Based on Intel Ivy-Bridge 12-core 2.7 GHz, with a total of 76,192 cores
- Nodes with 64 and 128 GB of main memory
- Network interconnect based on Cray Aries/Dragonfly
- Peak performance of 1.6 PFlop/s

Status of PRACE Services installation

Since network connectivity to EPCC has been setup a long time ago, only local configuration changes have to be made. These are currently in progress.

Accounting and data transfer services has been installed on ARCHER and are in production, as ARCHER is executing PRACE projects from DECI Calls. Other PRACE services are still pending installation - most of them depend on the network connectivity configuration.

5 Conclusions

The upgrade and integration procedure for new systems for both existing and new sites has been applied successfully to keep track of the evolution of the HPC resources devoted to PRACE and their integration. The procedure guarantees that sites plan well their upgrades and integration, and so minimizes the chance of problems in the transition period.
In a face to face meeting of WP6 in March 2014 the sites that applied this procedure were asked for feedback. All of them acknowledged the straight-forward use and organization of the procedure.