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 Phase Implementation Phase Project

Grant Agreement Number: RI-312763

D6.2.1
First Annual Integration Report

Final

Version: 1.0
Author(s): Gabriele Carteni, BSC, Jules Wolfrat, SURFsara
Date: 25.06.2013

Project and Deliverable Information Sheet

PRACE Project	**Project Ref. №: RI-312763**	
	Project Title: PRACE Third Phase Implementation Phase Project	
	Project Web Site: http://www.prace-project.eu	
	Deliverable ID: < **D6.2.1**>	
	Deliverable Nature: <DOC_TYPE: Report >	
	Deliverable Level: PU	**Contractual Date of Delivery:** 30/06/2013
		Actual Date of Delivery: 30/06/2013
	EC Project Officer: Leonardo Flores Añover	

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

Document Control Sheet

Document	**Title: First Annual Integration Report**	
	ID: D6.2.1	
	Version: 1.0	**Status: Final**
	Available at: http://www.prace-project.eu	
	Software Tool: Microsoft Word 2008 for Mac	
	File(s): D6.2.1.docx	
Authorship	**Written by:**	Gabriele Carteni, BSC Jules Wolfrat, SURFsara
	Contributors:	Marcello Morgotti, CINECA Gábor Rőczei, NIIFI George Tsouloupas, CaSToRC
	Reviewed by:	Dietmar Erwin, FZJ; Giacomo Mariani, CINECA
	Approved by:	MB/TB

Document Status Sheet

Version	Date	Status	Comments
0.1	10/04/2013	Draft	Index proposed
0.2	13/05/2013	Draft	Outline enhanced. Introduction and Chapter 2 edited.
0.3	29/05/2013	Draft	Added contributions for MareNostrum (BSC)
0.4	07/06/2013	Draft	Added contributions of NIIF and CaSToRC
0.5	11/06/2013	Draft	Added contribution for FERMI (CINECA).

			Draft formatted and finalised (Executive Summary, Annex, added). Ready for IR.
0.6	21/06/2013	Draft	Processed comments from Internal Review
1.0		Final version	Document consolidation

Document Keywords

Keywords:	PRACE, HPC, Research Infrastructure, Tier-0, PRACE Services, PFlop/s, BlueGene/Q, iDataPlex, SandyBridge

Table of Contents

List of Figures

List of Tables

References and Applicable Documents

[1] PRACE-2IP deliverable D6.3 Second Annual Operations Report of the Tier-1 Service
[2] ITIL v3 – Service Transition, http://www.itilv3.net/Service-Transition.html
[3] PRACE-3IP deliverable D6.1.1 First Annual Operations Report
[4] GEANT+ circuit service, https://prace-wiki.fz-juelich.de/pub/Prace2IP/WP6/IntegrationNIIF/Alternative_PRACE_connections_using_GEANT_services-v6.pdf

List of Acronyms and Abbreviations

AAA	Authorization, Authentication, Accounting.
AISBL	Association sans but lucrative (legal form of the PRACE RI)
AMD	Advanced Micro Devices
BSC	Barcelona Supercomputing Center (Spain)
CaSToRC	Computation-based Science and Technology Research Center, (Cyprus)
CINECA	Consorzio Interuniversitario, the largest Italian computing centre (Italy)
CSC	Finnish IT Centre for Science (Finland)
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
GPGPU	General Purpose GPU
GPU	Graphic Processing Unit
GSI-SSH	patched version of ssh which uses X.509 certificates to do authentication rather than traditional RSA keys or passwords. It is part of the Globus Toolkit
GRIDFTP	GridFTP is an extension of the standard File Transfer Protocol (FTP) for high-speed, reliable, and secure data transfer. It is part of the Globus Toolkit
HD	HelpDesk
HPC	High Performance Computing; Computing at a high performance level at any given time; often used synonym with Supercomputing
IB	InfiniBand
IEEE	Institute of Electrical and Electronic Engineers
INCA	Flexible framework to perform periodic, user-level functionality testing and performance measurement of Grid systems
IPERF	network testing tool that can create TCP and UDP data streams and measure the throughput of a network that is carrying them
IPSEC	Internet Protocol Security
LDAP	The Lightweight Directory Access Protocol
LINPACK	Software library for Linear Algebra
MPI	Message Passing Interface
NIIFI	National Information Infrastructure Development Institute (Hungary), also NIIF
PCIe	Peripheral Component Interconnect express, also PCI-Express
PCPE	PRACE Common Production Environment
PRACE	Partnership for Advanced Computing in Europe; Project Acronym
RAM	Random Access Memory
SURFsara	Dutch national High Performance Computing & e-Science Support Center
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
TTS	Trouble Ticket System
UNICORE	Uniform Interface to Computing Resources. Grid software for seamless access to distributed resources
XUUDB	UNICORE database providing a mapping from X.509 certificates to actual users' logins and roles

Executive Summary

This report describes the activities of Task 6.2 (WP6): Integration and upgrade of new Tier-0 and Tier-1 systems.

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 much 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 has been successfully applied for the upgrade of three systems: MareNostrum at BSC from Tier-1 to Tier-0, which is reported in this document, and Cartesius at SURFsara and SiSu at CSC. The upgrades of the latter two are both for Tier-1 systems and are managed by PRACE-2IP as part of the operations of the Tier-1 infrastructure [1]. The experience of these upgrades is used for improving the procedure itself. The upgrade procedure, as discussed in this document, consists of nine steps that should cover all possible scenarios.

Two new Tier-0 systems have successfully been taken in service in this project period, an IBM BlueGene/Q system at CINECA, Italy, and an IBM iDataPlex DX360M4 system at BSC, Spain. Both these systems replaced an existing Tier-1 system.

This task also started the integration of two new Tier-1 sites, NIIF in Hungary and CaSToRC on Cyprus. The procedure being used to integrate these systems is the same used by other Tier-1 systems in PRACE-2IP.

1 Introduction

The existing PRACE distributed Tier-0 and Tier-1 infrastructure is continuously extended by new Tier-0 and Tier-1 resources, as well as by upgrades of existing systems. In the first year the integration or upgrades of new Tier-0 systems was the responsibility of this task. The integration of two new Tier-1 sites also is the responsibility of this task.

This task defined a procedure for the upgrades of systems. The procedure and the experience with using it are described in section 2.

In section 3 the integration of two new Tier-0 systems is described, an IBM BlueGene/Q system at CINECA, Italy, and an IBM iDataPlex DX360M4 system at BSC, Spain.

The integration of two new Tier-1 sites, NIIF in Hungary and CaSToRC on Cyprus, has been started by this task and the status is given in section 4.

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 doesn't give a good overview of all changes involved with a system replacement and especially of all dependencies. This is not surprising because it was conceived for managing changes on services, not on systems. So a general guideline for the replacement of a system was developed, also because there will be many similarities between different system replacements. This procedure will help to improve the general availability and reliability of PRACE services.

After a preliminary study, which took into account local procedures as well as best practices [2], it was clear that a common process covering all possible scenarios was not viable due to many different conditions that may happen during the upgrade of a supercomputer in PRACE.

For example, if the new system physically replaces the old one, some actions cannot start before others. In this case, the installation process cannot start before a full or partial removal of the old system. On the other hand, if the new system can be installed on a different site, installation as well as acceptance tests can run independently from the decommission process of the old system. In some cases, both systems can run simultaneously and be integrated in PRACE, providing high levels of business continuity.

In order to address all different scenarios, the procedure that we are going to describe has been designed to allow flexibility and customization. It can be considered as a general and common guidance allowing partners to report their specific case. It has been designed for both Tier-0 and Tier-1 PRACE systems, since both are part of the PRACE infrastructure.

To ensure continuous service to users this procedure assumes that:

- The system being decommissioned has been withdrawn on time from any PRACE/DECI call and 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.

Most important requirement is that PRACE services are 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 Test
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

This procedure has been applied to monitor three system upgrades:

- **MareNostrum** at BSC (Spain), from Tier-1 to Tier-0
- **Cartesius** at SURFSara (The Netherlands), from Tier-1 to Tier-1
- **Sisu** at CSC (Finland), from Tier-1 to Tier-1

The two Tier-1 upgrades were the responsibility of PRACE-2IP, so are not further described in this report. These early use cases allow getting first feedback for tuning the procedure and making it more accurate and complete. The procedure is implemented by a live report internally published on the PRACE Wiki (section “Operations”) and constantly updated by the involved partner.

Changes on services are also tracked, as usual, through the PRACE Change Management Procedure, where all changes are considered as "Major". Impact on other sites must be considered, especially security implications, as well as impacts on central services, like Monitoring and User Documentation.

Step 1: Plan and Prepare the System Upgrade.

Good planning and management are essential. The impact of a system/service transition in a distributed environment, like PRACE, can produce high costs and expose the entire infrastructure to failures and/or a bad quality of service provisioning that can damage the image of the entire PRACE Research Infrastructure as well as that of single partners.

Every System Upgrade must be planned even though all needed information might not be initially available. A schedule is the result of this step and it can be further revised and improved during the process. It always serves as a reference document during the process because it identifies milestones, dependencies and timing. It also enables early risk detection on the operational infrastructure and possible mitigations. Organizational impacts can also be foreseen. Another benefit is that all partners are aware and up-to-date about any upcoming change. A template is available (see Annex 5.1).

Step 2: Internal Announcements.

All partners must be notified once the schedule of a new upgrade becomes available. This is accomplished through three actions:

- Adding an entry for the upcoming system in the System Information list that is regularly updated and maintained on the PRACE Wiki;
- Creating a specific Wiki Report for the upgrade process by inheriting and extending the provided template. Through this page all partners will stay up-to-date about the progress.
- Announcing the process to the Operations mailing list and the regular Operations video/phone-conferences held on a bi-weekly basis.

After this step, all partners are aware about a new upgrade process and it is expected that preliminary issues can be raised and addressed at this stage.

Step 3: Site Preparation.

Preparing a site to host a new machine is an activity that cannot find a place in a general and common timeline. It can start either very earlier before the procurement process or once the new system is well known in terms of site requirements (square meters, height of floating floors, building insulation degree, air flow capacity, etc…).

This step is not critical for the common operational activities of PRACE, so that only information like start and closing dates are expected in the Wiki Report.

Step 4: System Installation and Acceptance Test

Status of the installation of the new system as well as the result of the acceptance test should be reported in the Wiki Report and during the meetings (an announcement by mail is not needed). The completion of this step makes the new system ready for the integration in PRACE.

Step 5: Installation of PRACE services on the new system

A list of services that will be affected during the upgrade process must be provided. Among them, there are some that are supposed to be installed from scratch on the new system as replacement of the old ones (e.g. a new GSI-SSH login node, a new UNICORE resource, etc…).

For these services, outages can be expected if the old system cannot be kept online. Services that are loosely coupled to a specific system (e.g. Accounting Service, LDAP, etc…), should only suffer limited outages.

In order to assure a good level of business continuity and quality of service, all outages must be scheduled, dates must be provided and the Change Management procedure must be used.

Table 1 lists the set of core services, i.e. services that must be available on each PRACE system, along with a qualitative measure of the estimated effort for their installation. Effort estimation, described by three qualitative values (Low, Medium, High), is based on experience gained on managing services in PRACE. In any case, these estimations are intended as a general assessment since their computation doesn't consider any setback that might occur during an installation process.

Category / Service	Estimated Effort	Requirements
Network / Connection to the PRACE network	LOW	The new system must be connected to the local access point of the PRACE network.
Network / Monitoring	LOW	The network monitoring facilities must be migrated to the new system if they are hosted on the system that will be switched off.
Data / GridFTP	MEDIUM	If a separate system is available for bulk data transfers, then the only requirement is to mount the new user filesystem. If the storage is not part of the new system but an existing one is used, then this requirement is not valid anymore and outage is not expected. In case of a completely new installation is needed, then the setup could cost more than one person month, including tests and tuning.
Compute / Unicore	MEDIUM	At least the TSI component, responsible to interact with the local batch system, must be installed from scratch. Other components like Gateway, XUUDB and Unicore/X can live from the old system with minor modifications to the configuration files. In case a complete new installation is needed, then the setup could take about one person month, including tests and tuning.
AAA / LDAP	LOW	If the PRACE centralised LDAP service is used and the upgrade is from Tier-1 to Tier-0, then a new branch must be created and this usually takes a very short time. In case of an upgrade on the same tier (Tier-1 to Tier-1, or Tier-0 to Tier-0), then just an entry for the

Category / Service	Estimated Effort	Requirements
		new system should be added on an existing LDAP. If the LDAP service is provided locally by the site, then a new installation may be needed but it is not expected to take a long time in any case.
AAA / Accounting	MEDIUM	The accounting system relies on the local batch system, an internal database for storing accounting data and a web service for publishing it. Among these 3 components, only the local batch system should be part of the upgrade. So depending of the know-how available for the new batch system and the way it provides accounting information, more than 1 month can be allocated in general for restoring this service.
AAA / GSI-SSH	MEDIUM	The GSI-SSH service must be installed on a login node, one of the production login pools or a dedicated one. In any case, at least the "/home" filesystem and user accounts must be available as well as a link to the PRACE internal network.
User / Documentation	LOW	User Documentation requires rewriting at least information on how to get access, transfer data and submit jobs. Much of this information is limited to technical details like "hostname", "TCP port", "Software names", "Hardware details", etc… along with external links to site-specific documentation for further details. Effort is below one month.
User / PRACE HelpDesk	LOW	Setting up the central HD service for the new system is straightforward since it only requires adding a new "queue" and specific keywords to parse the subject of incoming mails for an automatic dispatching.
Monitoring / Inca	HIGH	Inca is a central and internal service responsible to monitor status of all PRACE services. For this reason, it should be the last action to be taken. A specific "reporter" must be installed locally for each service that has to be monitored. The installation of all reporters and a general test can take a significant amount of time, especially if specific customisations are needed.

Table 1: Expected service outages in a System Upgrade Process

Figure 1 shows a template available on the PRACE Wiki and used for keeping track of the different expected outages during the process for restoring the PRACE core services. It is intended to provide an up-to-date status of this critical step.

PRACE Services Transition

Service	Outage Planned?	Shutdown Date	Restore Date (Estimated)	Current Status
NETWORK / PRACE Link monitoring (iperf)	YES/NO	DD/MM/YYYY	DD/MM/YYYY	✓ / ✗
DATA / GridFTP	YES/NO	DD/MM/YYYY	DD/MM/YYYY	✓ / ✗
COMPUTE / UNICORE	YES/NO	DD/MM/YYYY	DD/MM/YYYY	✓ / ✗
AAA / LDAP	YES/NO	DD/MM/YYYY	DD/MM/YYYY	✓ / ✗
AAA / Accounting	YES/NO	DD/MM/YYYY	DD/MM/YYYY	✓ / ✗
AAA / GSISSH	YES/NO	DD/MM/YYYY	DD/MM/YYYY	✓ / ✗
USER / Documentation	YES/NO	DD/MM/YYYY	DD/MM/YYYY	✓ / ✗
USER / PCPE	YES/NO	DD/MM/YYYY	DD/MM/YYYY	✓ / ✗
USER / RT-TTS	YES/NO	DD/MM/YYYY	DD/MM/YYYY	✓ / ✗
MONITORING / INCA	YES/NO	DD/MM/YYYY	DD/MM/YYYY	✓ / ✗

Figure 1: Template used for monitoring the installation of PRACE services on a new system

Step 6: Pre-production

Once the new system enters into a pre-production phase, an announcement must be made through the three standard communication channels (Operation Mailing List, Wiki Report and Operation Meetings). Along with a formal communication, the partner running the upgrade should also report when they are ready to create the first accounts for PRACE users.

This phase usually is used for improving performance, stability and reliability.

Step 7: Shutdown of PRACE Services on the old system

Once all services have been restored, then the new system can be considered fully operational within the PRACE Research Infrastructure. The elapsed time between this step and Step#3 can be considered as the total duration of the Upgrade Process and further used for business analysis, calculating indexes like the business continuity rate, system transition time, etc.

The old system is then ready to be decommissioned or taken out of PRACE. First set of actions is shutting down the PRACE services on this system taking care about possible impacts on other sites (above all for script automations in place on other connected PRACE systems).

Step 8: Old System Shutdown

Before the shutdown of the old system, a further announcement must be sent through the 3 usual channels (Operation mailing list, Wiki Report and Operations Meetings).

Step 9: Review and Process closing

The last phase is intended for a last check that everything is working as planned.

During this step local staff, supported by selected PRACE partners, could start the PRACE Certification Process in order to check that all PRACE services have been correctly installed and configured. Certification for PRACE services is being consolidated as a procedure and it is basically an audit process periodically executed, in turn, by different PRACE partners.

3 Deployment of New Tier-0 Systems

3.1 FERMI – CINECA

The new CINECA Tier-0 system FERMI is a ten rack IBM BlueGene/Q system. Details about the configuration can be found in the deliverable D6.1.1 section 3 concerning the Tier-0 systems already in production [3]. The upgrade procedure, presented in chapter 2, was still not available when FERMI came into production (August 2012) and for this reason it was not followed.

The installation of FERMI went through different phases.

Since the best solution was to place FERMI in the same room where the old Tier-1 system, the IBM SP6 was placed in order to re-use as much as possible the water cooling system, the installation had to be planned largely in advance and a careful synchronization with all the operations necessary to SP6 decommissioning had to be considered.

An important preparation phase therefore started a long before the BlueGene/Q system was delivered to CINECA. This phase included all interactions with and re-allocation of users and related projects to other CINECA resources because of the shutdown of the SP6, a Tier-1 system, on May 17, 2012. The migration was to an IBM iDataPlex Tier-1 system, which also is used in PRACE, and an IBM BlueGene/P system where users could start to become familiar with the new Tier-0 architecture.

After these operations a lot of work started to adapt the power supply and especially the cooling system to the new system, which had different requirements in terms of water flow-rate and power output, in order to have everything ready for the delivery and setup of the machine which started at the beginning of July. In this phase took place also the setup of both the frontend and service nodes and of the 24 nodes I\O cluster with the respective Infiniband connectivity.

When the BlueGene/Q racks arrived in CINECA everything was ready for the hardware and software installation and the computing part of the system could be turned on quite quickly.

At the start of the FERMI operation in CINECA IBM has performed a sanity check of the hardware components and a basic setup of the system to perform and run preliminary Linpack. Later CINECA has gradually taken control of the machine performing other environmental tests and software benchmark to verify that the both the system and the I/O subsystem delivered the expected performance. At the end of July the configuration activity was completed with the necessary customizations to grant access to users. On August 1st, the machine was open in pre-production for the first PRACE users.

3.1.1 *Installation of PRACE Services*

Network / Connection to the PRACE network

FERMI was not connected to the PRACE 10Gbit Network initially, although a 10Gbit Internet connectivity was provided on all login nodes. The consequence was that CINECA door-node service was suspended for a while. At the beginning of 2013 two FERMI frontend nodes were connected to the PRACE network, restoring the door-node service.

Data / GridFTP

Public GridFTP was installed on the system as soon as it was declared "in production", i.e. September 2013. Once the PRACE link was connected also the PRACE GridFTP service was installed and enabled

Compute / UNICORE

Public CINECA Unicore Service was installed on the system as soon as the system was in production, PRACE users were automatically enabled to access the system.

AAA / LDAP

Since Tier-0 users and projects did not require to be defined in the PRACE LDAP system, but their creation was delegated to the Tier-0, CINECA relied on his own accounting system to enable these users.

Once the Door-node capability was restored also Tier-1 users registered in the PRACE LDAP were defined and enabled to access the frontend nodes to be able to access the PRACE GridFTP service over the PRACE dedicated network

AAA / GSISSH

Public GSI-SSH was installed on the system as soon as it was declared "in production", i.e. September 2013. Once the PRACE link was connected also the PRACE GSI-SSH service was installed and enabled

User / Documentation

User Documentation was made available for user from August 1st when the machine was opened to users. It was constantly improved and updated during the various upgrade of BlueGene/Q software stack, which took place until now.

User / PCPE

The PCPE was integrated with the CINECA module environment as soon as the system went in production.

User / RT-TTS

As for other Tier-0 systems the helpdesk for PRACE users was provided through local support and not through the PRACE helpdesk. However any request coming from the PRACE RT system concerning the Tier-0 system were attended to and resolved as quick as possible.

Monitoring / INCA

Consistent with its security policy CINECA has chosen to allow remote Monitoring from LRZ which is the site in charge of this activity. Thus, once the machine has been in production, INCA Monitoring was enabled and support was provided to LRZ colleagues for the setup.

3.2 MARENOSTRUM – BSC

The new Tier-0 system at BSC became available for user operations on 2nd of January 2013. Located in the same data centre as for the previous version of MareNostrum, which was a Tier-1 system in PRACE, it represents the third generation of the leading supercomputers at BSC. MareNostrum is based on Intel SandyBridge-EP Xeon E5-2670 processors, iDataPlex Compute Racks and Infiniband FDR-10, and has a peak performance of 1TFlop/s.

The upgrade process of MareNostrum needed a specific attention because it involved a physical replacement in the same computer room. The major consequence was that outage of some services like UNICORE, INCA and GSI-SSH could not be avoided.

During the site preparation, which required an improvement of cooling and power facilities, there was not any system available. In order to minimize any inconvenience, special attention was paid for not having any active project. This was accomplished by taking out the Tier-1

system from DECI calls in the an allocation period between September 2012 and December 2012 (time window for the site preparation and Tier-0 installation).

3.2.1 *System Upgrade*

The System Upgrade process of MareNostrum was different from the general one described in Chapter 2, due to site-specific conditions as the shared data centre between old and new system. According to the defined procedure, deviations have been applied as follows:

- Step 6 "Pre-production" added to Step 4 "System Installation and Acceptance Test"
- Step 7 "Shutdown of PRACE Services on the old system" added to Step 2 "Internal Announcements"
- Step 8 "Old System Shutdown" added to Step 2 "Internal Announcements"

Titles of Step 2 and Step 4 have been modified to reflect these changes and the resulting process was carried out as follows:

- **Step 1**: **Plan and Prepare the System Upgrade**
- **Step 2: Internal Announcements, PRACE Services and Old System Shutdown**
- **Step 3: Site Preparation**
- **Step 4: System Installation, Acceptance and Pre-Production Tests**
- **Step 5: Installation of PRACE services on the new system**
- **Step 6: Review and Process closing**

Step 1: Plan and Prepare the System Upgrade

The schedule for the upgrade process of MareNostrum was prepared by keeping input from all stakeholders participating to this activity (systems vendors, facilities vendors, system integrators). It was carried out during summer 2012, after the conclusion of the procurement process.

Step 2: Internal Announcements, PRACE Services and Old System Shutdown

Announcements to PRACE partners were made through regular videoconferences of WP6 and by updating a specific wiki page as well as the System Information list, also maintained in the PRACE wiki. The old system was shutdown on February 17th after a long period of periodic warnings to users started on July. It is worth noting that user data (included archived one) always remained available.

Step 3: Site Preparation

Site preparation was the longer part of the process due to the removal of the old system and the adaptation of the computer room in order to provide technical requirements for the new machine, like the new water circuit for cooling racks.

Figure 2: Pictures taken during the site preparation

Improvements mainly targeted cooling and power facilities. Power facilities were upgraded with two distribution transformers from 1000KVA to 2000KVA achieving a total of 5000KVA, which represented 70% of increase on the total power capacity. A new cabling layout was also needed in the data centre "The Chapel".

Two new chillers were also installed in a noise-reduction building for doubling the previous cooling capacity and integrated with existing five chillers. The total cooling capacity increased to 2202 kW.

Site preparation started in September 2012, after the shutdown of the old system and finally closed on early December.

Step 4: System Installation, Acceptance and Pre-Production Tests

System placement started as soon as the computer room was able to host the new system.

This happened during the second half of November in parallel with the test phase for facilities.

In mid of December systems was wired, powered, configured with all system software stack installed. The last 2 weeks of the year 2012 were devoted to carry out preliminary tests and first tunings.

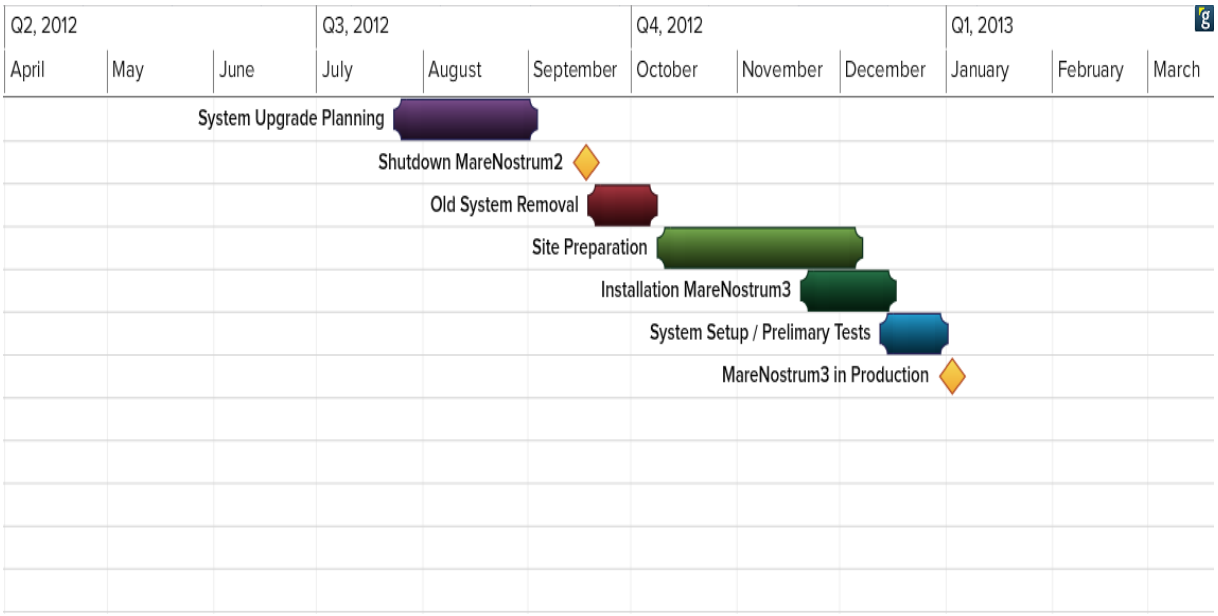


Figure 3: Gantt chart from Step1 to Step4

MareNostrum3 entered in production on January 2nd, 2013.

Step 5: Installation of PRACE services on the new system

At the time of this writing, installation of PRACE services is not yet completed. Detailed information is described in the next section.

Step 6: Review and Process closing

The upgrade process is not finished so that this step has not started yet.

3.2.2 *Installation/Restore of PRACE Services*

Network / Connection to the PRACE network

The link to the PRACE network has never been affected. BSC has always been connected to the PRACE network during the upgrade process and network monitoring has never stopped. The last network path from the gateway to the PRACE network located at BSC and the new system is related with services like GridFTP and GSI-SSH, which are currently under installation.

Data / GridFTP

GridFTP is not yet available. This service has the higher priority due to the probability to be requested by users and it is planned to have it installed by the end of June 2013.

It will be installed on a machine dedicated to data transfers, already available and able to provide resources (memory, bandwidth for I/O operations, network) for an high performance bulk data transfer.

Compute / UNICORE

UNICORE is not yet available for MareNostrum. Plan is to install it after GridFTP and GSISSH services.

AAA / LDAP

LDAP branch for MareNostrum is operated by the central PRACE LDAP service and available since the installation of MareNostrum. First users and projects were correctly added.

AAA / GSISSH

Along with GridFTP service, GSISSH is planned to be available by the end of June 2013. It is worth mentioning that external access to MareNostrum is always available via SSH with key exchange.

User / Documentation

User documentation is up-to-date to the PRACE website. Documentation about specific PRACE services will be updated service by service during their installation.

User / PCPE

The Module environment has been installed following the internal PRACE specifications.

User / RT-TTS

Queue for MareNostrum is correctly available in the central PRACE TTS. User support is provided through the BSC TTS and the mail prace-support@bsc.es is used.

Monitoring / INCA

INCA service will be the last service to be installed since it is responsible to monitor the status and health of all PRACE services. It is expected to have MareNostrum monitored by INCA from September 2013. This date is also the start of the last step of the System Upgrade Process (Step 6: Review and Process closing).

4 New Tier-1 Systems

Two Tier-1 systems were scheduled for integration by this task. They are hosted and operated by NIIFI (Hungary) and CaSToRC (Cyprus). Ongoing activities, currently managed by WP6 in PRACE-2IP, related to the integration of other Tier-1 systems will become the responsibility of PRACE-3IP in the second year after the end of PRACE-2IP.

In this chapter we present the activities undertaken with NIIFI and CaSToRC, such as assignment of supporting partner, network planning, and work plans for service installation. All these activities are needed for a timely integration of these partners before the start of DECI-11 projects in November 2013.

4.1 NIIFI

Hardware information

The offered NIIFI supercomputer is a fat-node HP cluster which is a very sophisticated type of blade technology (CP4000BL). It is using the latest AMD Opteron 6174 type processors with 12-core Magny-Cours. Total number of cores is 768. The interconnect network is a redundant QDR Infiniband network.

The whole system has two TBytes of memory and the computing power is 5.4 TFlop/s. Water-cooled racks are used in the system to increase the energy efficiency. This unique supercomputer runs very effectively in the mixed parallel programming paradigms and each node is a powerful 24 cores SMP computer.

First contacts and activities

The following actions were taken as part of the integration:

- Access to the internal services (Wiki, TTS, BSCW, SVN)
- Proposed an alternative solution instead of IPsec to connect new members via GEANT+ circuit service [4].
- GSI-SSH and GridFTP servers have been installed; LDAP and grid-mapfile integration completed
- UNICORE GW, TSI, UNICORE/X, XUUDB installed on supercomputer; LDAP and XUUDB integration completed
- LDAP-branch successfully accessed, populated with staff members, and associated groups modified. Integration with NIIFI LDAP completed
- PCPE environment and prace-service script have been installed
- INCA monitoring installed
- Draft of the user documentation completed

4.2 CaSToRC

Hardware information

Cy-Tera is the first National HPC resource is Cyprus. Commissioned in early 2012, Cy-Tera is a 35TFlop/s CPU/GPU Hybrid machine. In more detail, Cy-Tera features the following:

- 116 twelve-core iDataPlex dx360 M3 compute nodes, 18 of which with NVidia GPUs
- Two 6-core CPU packages (Intel Xeon X5650) per node
- 36 NVidia M2070 GPU
- 48 GB memory per node

- Total memory 4.7 TBytes
- 4x QDR Infiniband network for MPI and for I/O to the global GPFS filesystem
- 300TBytes of storage
- Measured storage access rate at 4.7GBytes/s
- 160TByte Long-term storage

First contacts and activities

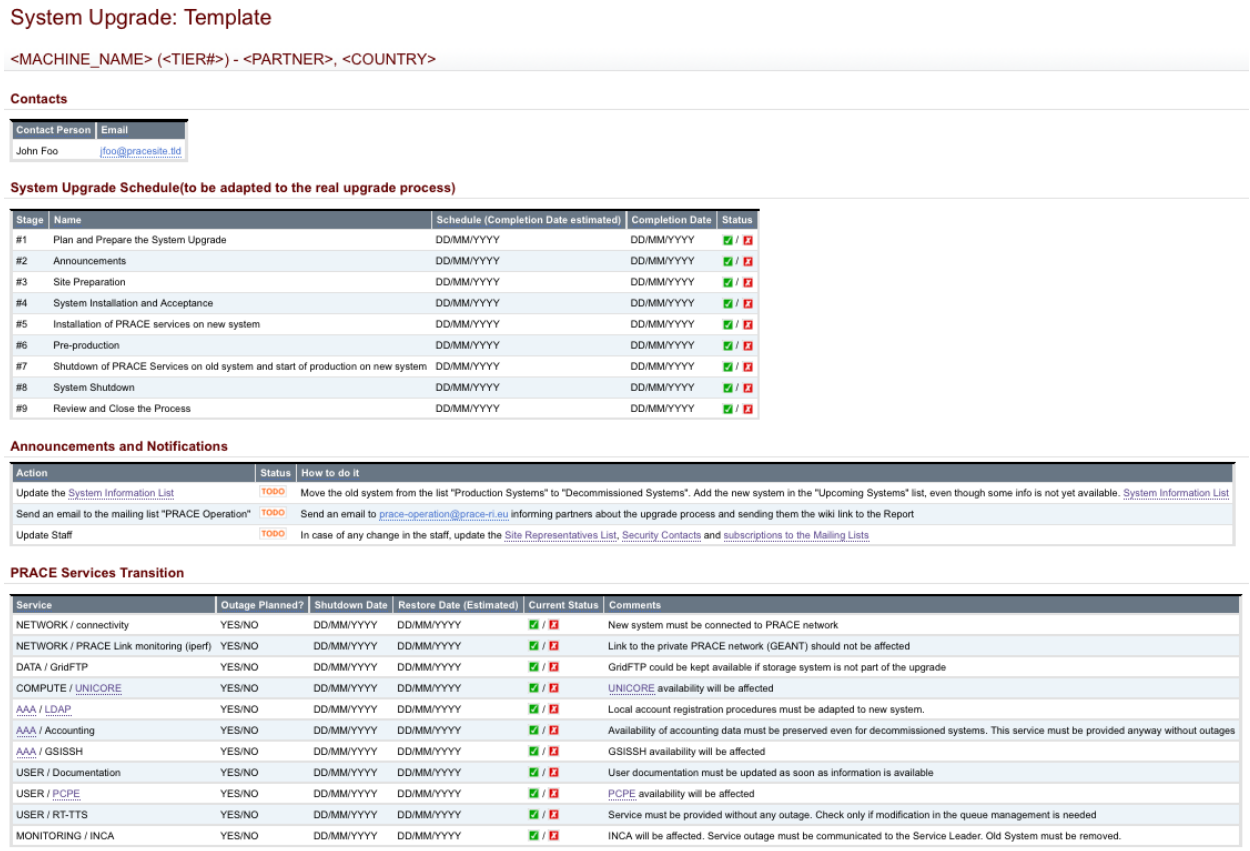
Contact points and the team that supports the system have been communicated. VPN equipment to be connected through IPsec to the PRACE Network is in place and ready to be configured.

Hardware to host and operate PRACE services (e.g. UNICORE, GridFTP, GSI-SSH) has been identified and allocated.

5 Annexes

5.1 Annex 1: Template for a System Upgrade

A template implementing the procedure for a system upgrade documented on Chapter 2 is available on the PRACE Wiki. The following figure is a snapshot taken from a web browser.



System Upgrade: Template

<MACHINE_NAME> (<TIER#>) - <PARTNER>, <COUNTRY>

Contacts

Contact Person	Email
John Foo	jfoo@pracesite.tld

System Upgrade Schedule(to be adapted to the real upgrade process)

Stage	Name	Schedule (Completion Date estimated)	Completion Date	Status
#1	Plan and Prepare the System Upgrade	DD/MM/YYYY	DD/MM/YYYY	✓ / ✗
#2	Announcements	DD/MM/YYYY	DD/MM/YYYY	✓ / ✗
#3	Site Preparation	DD/MM/YYYY	DD/MM/YYYY	✓ / ✗
#4	System Installation and Acceptance	DD/MM/YYYY	DD/MM/YYYY	✓ / ✗
#5	Installation of PRACE services on new system	DD/MM/YYYY	DD/MM/YYYY	✓ / ✗
#6	Pre-production	DD/MM/YYYY	DD/MM/YYYY	✓ / ✗
#7	Shutdown of PRACE Services on old system and start of production on new system	DD/MM/YYYY	DD/MM/YYYY	✓ / ✗
#8	System Shutdown	DD/MM/YYYY	DD/MM/YYYY	✓ / ✗
#9	Review and Close the Process	DD/MM/YYYY	DD/MM/YYYY	✓ / ✗

Announcements and Notifications

Action	Status	How to do it
Update the System Information List	TODO	Move the old system from the list "Production Systems" to "Decommissioned Systems". Add the new system in the "Upcoming Systems" list, even though some info is not yet available. System Information List
Send an email to the mailing list "PRACE Operation"	TODO	Send an email to prace-operation@prace-ri.eu informing partners about the upgrade process and sending them the wiki link to the Report
Update Staff	TODO	In case of any change in the staff, update the Site Representatives List, Security Contacts and subscriptions to the Mailing Lists

PRACE Services Transition

Service	Outage Planned?	Shutdown Date	Restore Date (Estimated)	Current Status	Comments
NETWORK / connectivity	YES/NO	DD/MM/YYYY	DD/MM/YYYY	✓ / ✗	New system must be connected to PRACE network
NETWORK / PRACE Link monitoring (iperf)	YES/NO	DD/MM/YYYY	DD/MM/YYYY	✓ / ✗	Link to the private PRACE network (GEANT) should not be affected
DATA / GridFTP	YES/NO	DD/MM/YYYY	DD/MM/YYYY	✓ / ✗	GridFTP could be kept available if storage system is not part of the upgrade
COMPUTE / UNICORE	YES/NO	DD/MM/YYYY	DD/MM/YYYY	✓ / ✗	UNICORE availability will be affected
AAA / LDAP	YES/NO	DD/MM/YYYY	DD/MM/YYYY	✓ / ✗	Local account registration procedures must be adapted to new system.
AAA / Accounting	YES/NO	DD/MM/YYYY	DD/MM/YYYY	✓ / ✗	Availability of accounting data must be preserved even for decommissioned systems. This service must be provided anyway without outages
AAA / GSISSH	YES/NO	DD/MM/YYYY	DD/MM/YYYY	✓ / ✗	GSISSH availability will be affected
USER / Documentation	YES/NO	DD/MM/YYYY	DD/MM/YYYY	✓ / ✗	User documentation must be updated as soon as information is available
USER / PCPE	YES/NO	DD/MM/YYYY	DD/MM/YYYY	✓ / ✗	PCPE availability will be affected
USER / RT-TTS	YES/NO	DD/MM/YYYY	DD/MM/YYYY	✓ / ✗	Service must be provided without any outage. Check only if modification in the queue management is needed
MONITORING / INCA	YES/NO	DD/MM/YYYY	DD/MM/YYYY	✓ / ✗	INCA will be affected. Service outage must be communicated to the Service Leader. Old System must be removed.

Figure 4: Snapshot from the PRACE Wiki of the Template for a System Upgrade

5.2 Annex 2: Tier-0 List and Integration Year

System	Architecture	Partner	Support through	Start Date
Curie (Fat)	Bull Bullx BCS	CEA	PRACE-3IP	2011
Curie (Hybrid)	Bull Bullx B505	CEA	PRACE-3IP	2011
Curie (Thin)	Bull Bullx B510	CEA	PRACE-3IP	2011
FERMI	IBM BlueGene/Q	CINECA	PRACE-3IP	2012
Hermit	Cray XE6	HLRS	PRACE-3IP	2011
JUQUEEN	IBM BlueGene/Q	FZJ	PRACE-3IP	2012
MARENOSTRUM	IBM iDataPlex DX360M4	BSC	PRACE-3IP	2013
SuperMUC	IBM iDataPlex DX360M4	LRZ	PRACE-3IP	2012