SEVENTH FRAMEWORK PROGRAMME
Research Infrastructures

INFRA-2010-2.3.1 – First Implementation Phase of the European High Performance Computing (HPC) service PRACE

PRACE-1IP

PRACE First Implementation Project

Grant Agreement Number: RI-261557

D5.2.2
PRACE Industrial Offer Assessment
Final

Version: 1.2
Author(s): Stephane Requena, GENCI
Date: 24.05.2012

Project and Deliverable Information Sheet

PRACE Project	**Project Ref. №: RI-261557**	
	Project Title: PRACE First Implementation Project	
	Project Web Site: http://www.prace-project.eu	
	Deliverable ID: D5.2.2	
	Deliverable Nature: REPORT	
	Deliverable Level: PU *	**Contractual Date of Delivery:** 31 / 05 / 2012
		Actual Date of Delivery: 31/ 05/ 2012
	EC Project Officer: Thomas Reibe	

* - 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: PRACE Industrial Offer Assessment**	
	ID: D5.2.2	
	Version: 1.2	**Status:** Final
	Available at: http://www.prace-project.eu	
	Software Tool: Microsoft Word 2011	
	File(s): D5.2.2-PublicVersion.docx	
Authorship	**Written by:**	Stephane Requena, GENCI
	Contributors:	Marcin Ostasz, BSC Carlos Merida, BSC Nicolas Mignerey, GENCI
	Reviewed by:	Florian Berberich, FZJ Mark Parsons, EPCC
	Approved by:	MB/TB

Document Status Sheet

Version	Date	Status	Comments
0.1	20/04/2012	Draft	Initial Version
0.2	04/05/2012	Draft	New version with add-dons from Marcin and Nicolas
0.3	10/05/2012	Draft	Section on the Bull offer and little polishing by Nicolas
0.4	11/5/2012	Draft	Inclusion of legal assessment in the annexes
1.0	14/05/2012	Ready for internal review	

Version	Date	Status	Comments
1.1	21/05/2012	Draft	Incorporation of the comments for reviewers.
1.2	24/05/2012	Final	Final version with last comments

Document Keywords

Keywords:	PRACE, HPC, Research Infrastructure, Industrial Offer, business models, services

Table of Contents

List of Figures

List of Tables

List of Acronyms and Abbreviations

AWS	Amazon Web Services
BSC	Barcelona Supercomputing Centre (Spain)
CCRT	Centre Commun de Recherche et Technologie (France)
CEA	Commissariat à l'Energie Atomique et aux Energies Alternatives (represented in PRACE by GENCI, France)
CINECA	Consorzio Interuniversitario, the largest Italian computing centre (Italy)
CINES	Centre Informatique National de l'Enseignement Supérieur (represented in PRACE by GENCI, France)
DoE	US Department of Energy
EC	European Commission
EPCC	Edinburgh Parallel Computing Centre (represented in PRACE by EPSRC, United Kingdom)
EPSRC	The Engineering and Physical Sciences Research Council (United Kingdom)
ETP	European Technology Platform
ETHZ	Eidgenössische Technische Hochschule Zuerich, ETH Zurich (Switzerland)
ESFRI	European Strategy Forum on Research Infrastructures; created roadmap for pan-European Research Infrastructure.
GCS	Gauss Centre for Supercomputing, alliance of the three national german supercomputer centres FZJ-JSC, USTUTT-HLRS and BADW-LRZ
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)
HLRS	Höchstleistungsrechenzentrum Stuttgart represented in PRACE by GCS
HPC	High Performance Computing; Computing at a high performance level at any given time; often used synonym with Supercomputing
HPL	High Performance LINPACK
HQ	Headquarters

IDRIS	Institut du Développement et des Ressources en Informatique Scientifique (represented in PRACE by GENCI, France)
INCITE	Innovative and Novel Computational Impact on Theory and Experiment
ISV	Independent Software Vendor
IT	Information Technology
NCF	Netherlands Computing Facilities (Netherlands)
NREN	National Research and Education Network
PaaS	Platform as a Service
PATC	PRACE Advanced Training Centers
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
RI	Research Infrastructure
ROI	Return of investment
SC'11	Supercomputing 2011 conference, held in Seattle on November 2011
SaaS	Software as a Service
SME	Small and Medium size Enterprise
SWOT	Strengths, Weaknesses, Opportunities and Threats
WP2	PRACE First Implementation Phase Work Package 2: "Evolution of the Research Infrastructure »
WP4	PRACE First Implementation Phase Work Package 4: "HPC Ecosystem Relations »
WP5	PRACE First Implementation Phase Work Package 5: "Relations with industrial users"
WP9	PRACE Second Implementation Phase Work Package 9: "Industrial applications support"

Executive Summary

This document contains the result of the assessment conducted by WP5 of the different business models and services based on the expectations of industrial users and how adequate these requirements are with regard to technical, organisational, legal and financial aspects relative to the PRACE Research Infrastructure.

Three business models: Open R&D, Pre Competitive R&D and Production/Commercial activities with high value services like training, code enabling/optimisation, access to HPC resources, technology watch and co-development have been investigated by WP5 in collaboration with WP2.

The outcome of this work will be used by the PRACE Council in making its decision on how to support a new business model and its associated services to industry through PRACE. Based on a preliminary version of this document, the PRACE Council has already taken a decision on January 24nd of 2012 to open existing PRACE calls for proposals to industry on the basis of Open R&D projects.

Starting from the 5th PRACE regular call for proposals, industrial users may apply to PRACE resources and services as Principal Investigator, alone or together with academia. The projects will be selected purely on the basis of scientific excellence as it is already done for projects from academia.

This decision is opening a door to a wide area of collaboration and technology transfer between academia and industry and will foster the use of HPC for improving European industrial competitiveness and innovation.

1 Introduction

The major goal of this document is to deliver the results of the assessment of the potential business models and services that PRACE may provide to commercial companies as end users of its facilities.

This work has been conducted by work package WP5 after an extensive analysis of existing business models and services performed in D5.2.1 “Business models for industrial access to infrastructures” and the gathering of industrial users realised into D5.1.2 “Requirements of industrial users”.

Section 2 presents the results of the needs and expectations gathering exercise, based on the outcome of two industrial seminars, face-to-face meetings and surveys organised by WP5 during the period. These results are complemented by the outcome of six different industrial pilots with different types of companies, services and locations in order to assess the impact of enrolling such pilots into the PRACE Research Infrastructure in terms of technical and organisational issues.

Section 3 describes the different industrial target groups that have been identified for PRACE, and which high value services could be developed in order to reach user expectations.

Finally, Section 4 presents the assessment of the three different business models against technical, legal and financial constraints relative to the status of the PRACE Research Infrastructure as well as the four PRACE Hosting Members (France, Germany, Italy and Spain).

2 Requirements of industrial users

Following the outcome of the previous PRACE Preparatory Program project regarding the establishment of a link with potential industrial users of the PRACE RI, PRACE 1IP WP5 worked on finalising the capture of the requirements of industrial users.

This activity was performed through a four-step approach:

- By organising two industrial seminars in 2011 in Stockholm and 2012 in Bologna
- By setting up, as a follow up to such seminars, face to face meetings with companies interested in using forthcoming PRACE services for industry
- By launching, in collaboration with PRACE-2IP WP9, surveys about "general" industrial users and ISV needs
- By enrolling pilots during the period of the PRACE-1IP project in order to evaluate the impact on the infrastructure and gather their feedback about their first usage of the facilities and the services.

In D5.1.2 (Dec 2011) WP5 summarised the needs of the industrial users. The current HPC industrial ecosystem can be divided into the following categories of organisations, according to their needs:

- Desktop-level SMEs – companies of less than 250 employees that have not yet adopted HPC
- SME-level HPC users – companies of less than 250 employees that are using HPC (and intend to extend its usage)
- Independent Software Vendors – software companies providing independent application software
- Large organisations – companies of more than 250 employees

The needs of the four groups are as follows:

- Evangelisation, information and training – access to expertise on the possibilities of HPC as well its implementation and maintenance requirements;
- On-demand access to HPC resources which provides the opportunity to 'try-out' HPC solutions (without committing extensive resources);
- Expertise in scientific software by supporting and recommending ISV code options as well as creating a pool of Open Source Codes - in order to allow smaller users to adopt HPC;
- Code Enabling – the scaling of codes up to Tier-0 level;
- Co-design of industrial applications – cooperation with PRACE in order to develop new software solutions for industry;
- An Open R&D Programme in order to perform R&D and develop solutions that can be shared with other members of community;
- Pre-Competitive R&D - R&D activity or code development where the disclosure of results may be postponed or which might require a payment;
- Partnership with HPC research centres in order to develop solutions and participate in PRACE prototyping activities;
- Technology Transfer and other high-value knowledge services – access to knowledge on HPC hardware and software technology options;
- Commercial Activities – access to resources on a commercial basis.

The matrix below illustrated the groups of users that should be targeted with the various initiatives outlined above.

Various combinations of the services above will meet the requirements of different user groups. The table below illustrates a number of examples of programmes that might be developed in order to offer such services to various user groups. The grey areas indicate the areas where a specific requirement has been detected.

	User Group			
Requirement	Desktop-level SMEs	SME-level HPC users	Independent Software Vendors	Large organisations
Evangelisation				
On-demand access				
Expertise in scientific software				
Code Enabling				
Co-design of industrial applications				
An Open R&D Programme				
Pre-Competitive R&D				
Partnership with HPC Research Centres				
Technology Transfer				
Commercial Activities				

Table 1 : Industrial User Groups and their Requirements (the grey areas indicate the areas where a specific requirement has been detected).

2.1 PRACE 1IP industrial seminars

As a continuation of the activities of the PRACE PP WP3, and during the subsequent period, PRACE-1IP WP5 organised two industrial seminars, one in Stockholm in April 2011 and one in Bologna in April 2012. These events were held with the goal of creating strong links with industrial end-users, informing them about the deployment of PRACE facilities and services, determining their needs and expectations as well as promoting the use of HPC in industry as a tool for increasing industrial competitiveness.

In total, the four industrial seminars organised by PRACE have gathered 373 executive attendees representing around 120 different companies from various industrial domains such as

Energy, Materials, Aeronautics, Materials, Automotive, Life Sciences/Pharma as well as Digital Media, Finance, Renewable Energy and Micro Electronics.

One of the key drivers of the seminars was to be able to attract new industrial domains, to stimulate attendance from many European countries and to expand the awareness of HPC to SMEs. For example, the third industrial seminar in Stockholm attracted attendees from more than 22 countries. Most recently, 42% of the attendees of the fourth event in Bologna were from SMEs. For more information about the organisation and the results of the two seminars, two documents were issued by WP5: D5.1.1 “First industrial seminar » and D5.1.3 « Second industrial seminar ».

Beyond information about the deployment of the PRACE infrastructure and its services and beyond the networking between attendees, such events have been an opportunity to discuss, in parallel sessions with the attendees, which business models to propose, which services to offer, how to involve SMEs, which relations are most appropriate ISVs and the Open Source communities, …

One of the major pieces of feedback from these discussions was that industrial users consider PRACE as an enabler to drive company adoption of HPC and to foster technology transfer between academia and industry.

During these four industrial seminars, attendees have been able to assess the progressive development of the PRACE industrial offer. The fourth seminar was a unique opportunity for PRACE AISBL to announce the launch of the Open R&D offer for industry.

2.2 Face to face meetings

In the interval between the industrial seminars, PRACE-1IP WP5 and previously PRACE PP-WP3 set up dedicated face-to-face meetings with industrial users. These meetings provided an opportunity to address specific needs and to discuss more transparently the expectations of such companies regarding PRACE. Larges companies as well as SMEs were interviewed, covering a wide range of industrial domains from Energy, Finance, Materials, Automotive, Service and ISVs.

Similar to the feedback from the industrial seminars, these industrial users expect PRACE to be an enabler that allows companies to adopt HPC and to foster technology transfer between academia and industry. Some large companies already have their own HPC facilities but these are reserved for “production” studies. Such companies want to develop access for their R&D department to PRACE facilities in order to strengthen the link with the academic research on non-confidential data. For “production” studies, the level of confidentiality of the data is generally too high to allow the outsourcing of such activity to PRACE facilities or even to HPC Cloud providers. Such companies are much more interested in deploying their own internal Cloud services and sharing them with their associated supply chain.

For ISVs PRACE represents an interesting dissemination opportunity in order to address a wide range of new users from academia and industry among the 24 member states. However, issues such as better licensing costs or the increase of the parallel performance of ISV codes needs to be addressed in order to reach such goals. ISVs also showed an interest in being involved in PRACE training activities like the ones provided by the PATC (PRACE Advanced Traning Centers) or the PRACE Online Training Portal.

For all the companies, PRACE is seen as a provider of not only CPU cycles, but also of high value services (information, training, code enabling, ISV and open source promotion, technology watch, prototyping, best practice, etc.,), due to the cumulative HPC experience of its twenty-four partners.

2.3 PRACE-1IP and -2IP surveys

PRACE-1IP deliverable D5.1.2 titled 'The Requirements of Industrial Users', submitted in December 2011, analysed the needs of various types of industrial organisations and provided an overview of potential models that might satisfy those needs. The analysis involved a review of other sources available in Europe and in other regions as well as the results of an ongoing survey launched for the purpose of the task.

To date, 34 companies have responded to the survey. It targeted the attendees of the PRACE Industrial Seminars. The survey has identified three major obstacles faced by companies intending to either begin using HPC or extending its usage. They are:

- The cost of implementation;
- Lack of appropriate, easy-to-use software;
- Access to expertise (HPC, scientific expertise, …)

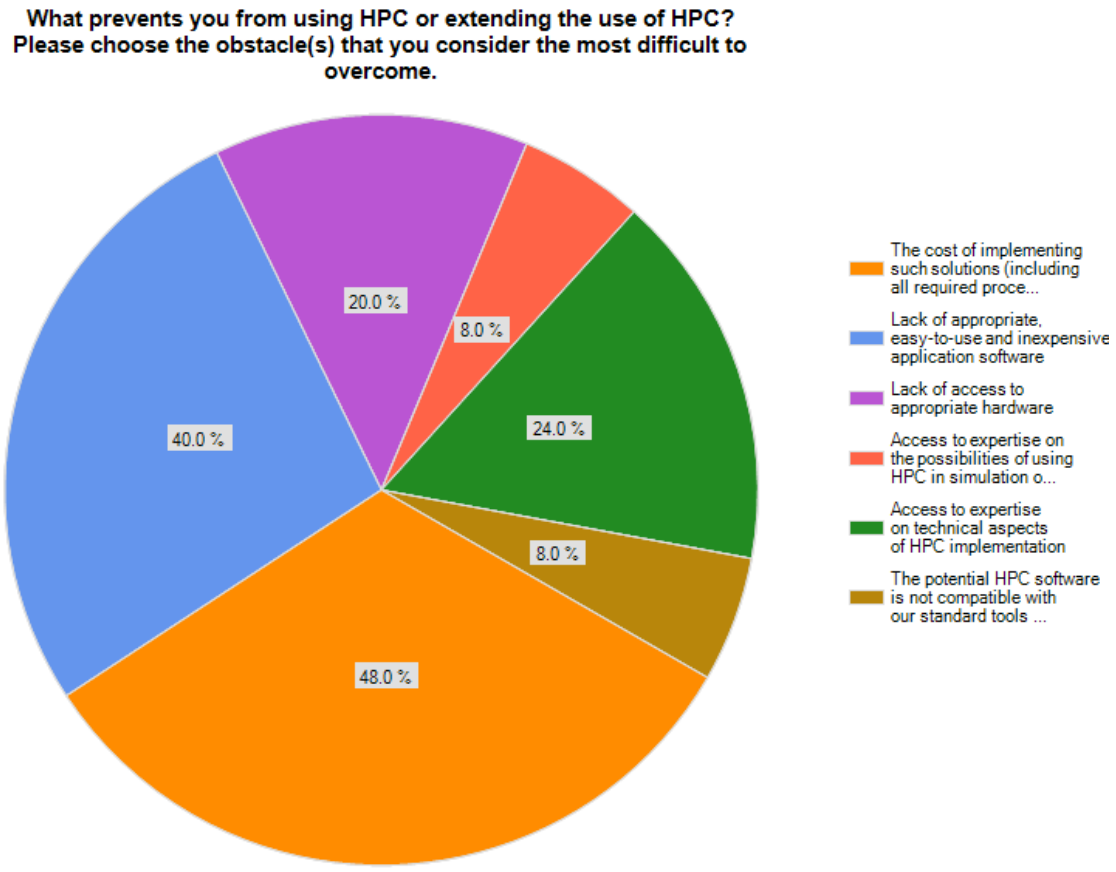

Figure 1 – Survey result 1: Main obstacles to overcome for using HPC resources

The major factors that would motivate companies to adopt or broaden the use of HPC are:

- Partnership programmes with research organisations or other companies
- Availability of open-source application software
- On-demand access to HPC resources

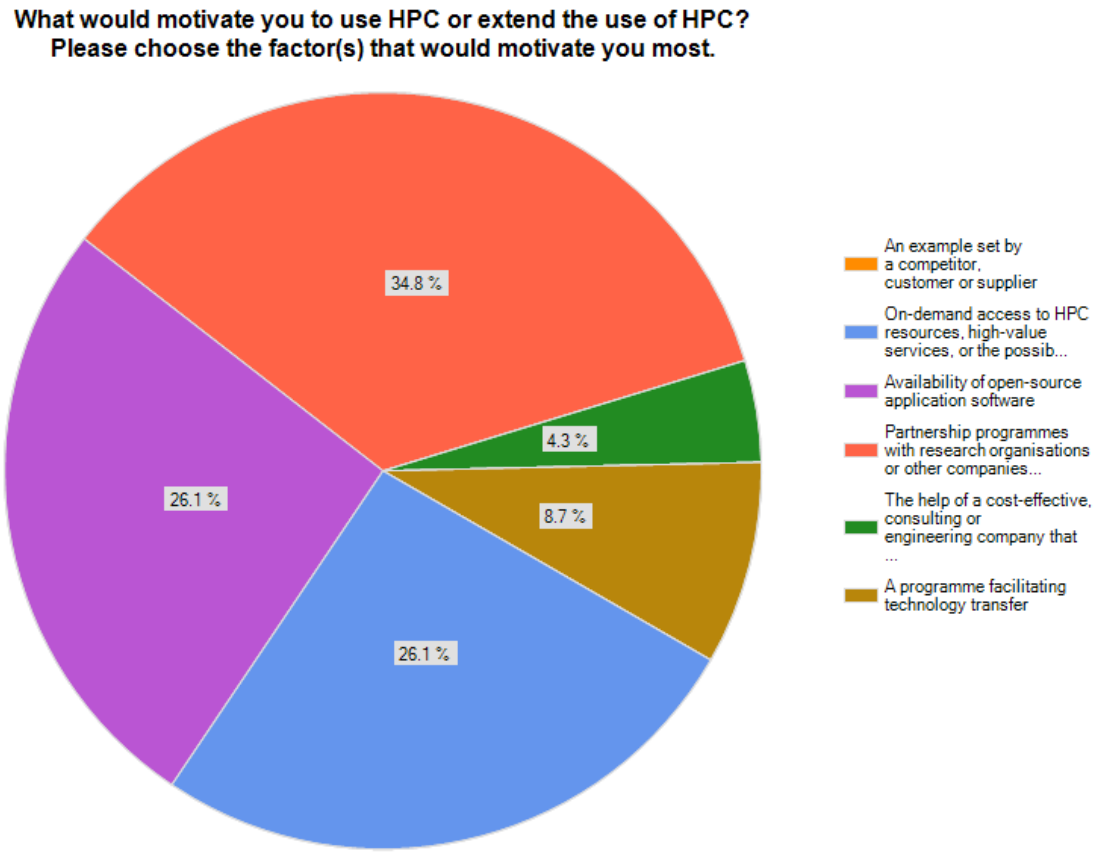

Figure 2 - Survey result 2: Motivations for using HPC

Regarding the first possible answer in the survery there were no respondents who would have selected the first option: An example set by competitior.

When asked what business model should be adopted by PRACE in order to help address the obstacles identified above, the respondents identified the following options:

- Open R&D
- Pre-Competitive R&D
- Commercial activities

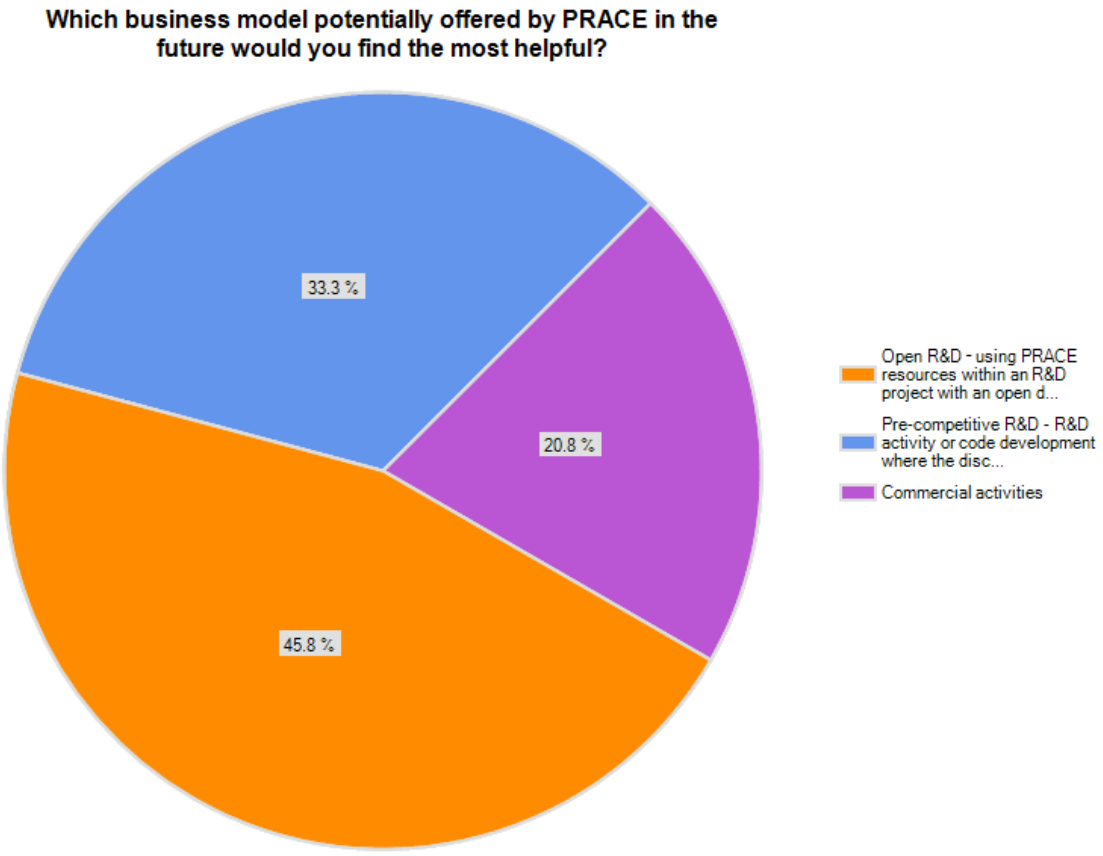

Figure 3 - Survey result 3: Which business models will be helpful for industry?

On a more granular level, such a business model should allow the industrial users to take advantage of the following services, as identified in the survey:

- Access to HPC Cycles
- Technology Watch
- Code enabling
- Prototyping

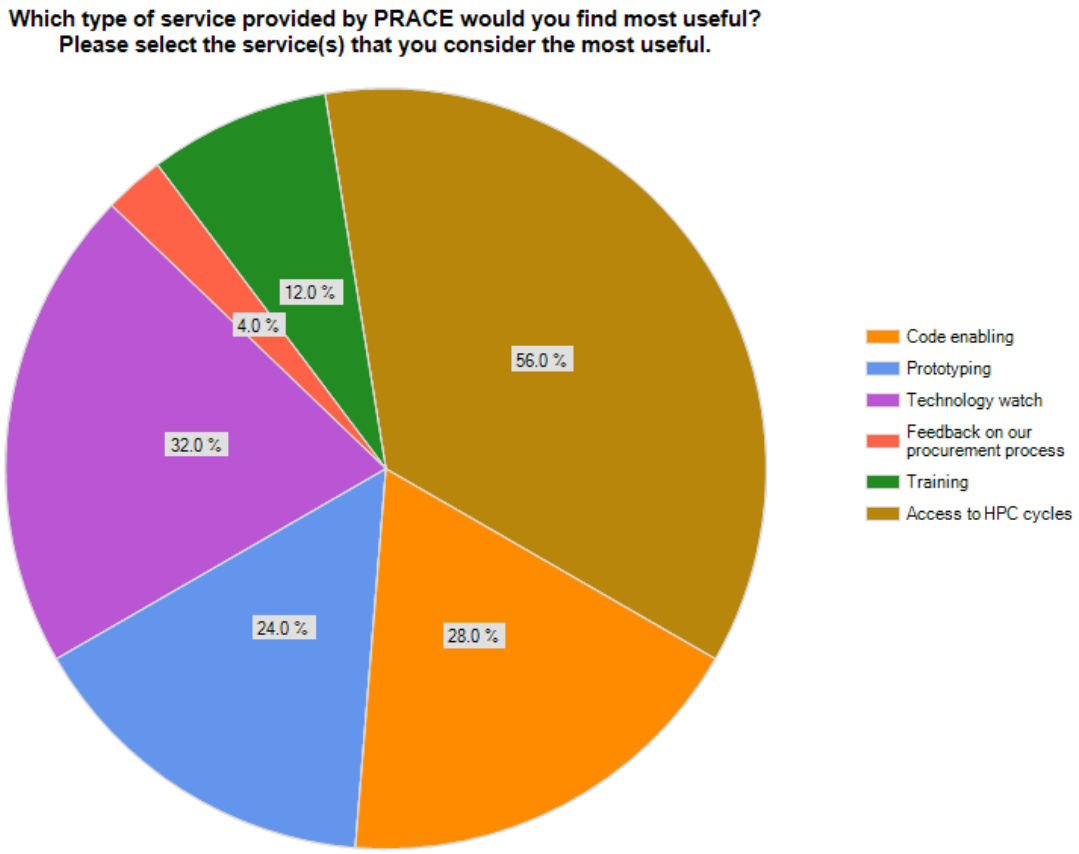

Figure 4 - Survey result 4: Which high value services are the most useful?

2.4 Open R&D industrial pilots

In order to improve the understanding of the needs and expectations of industrial users and to send an initial positive signal to industry about the desire of PRACE to be able to deliver a number of services to industry, a set of industrial pilots have been enrolled into the PRACE RI during the PRACE-1IP project by WP5. These pilots, proposed by the PRACE partners involved into WP5 activities, have been selected in order to cover initially:

- A wide number of industrial domains;
- A wide number of European countries;
- A wide range of companies, from large companies to SMEs and ISVs;
- The list of potential services and business models which may be proposed by the AISBL to industry;
- A wide range of HPC architectures.

In addition, to improve the understanding of the needs and expectations of industry, such pilots have been used for assessing the impact of enrolling industrial users on the operational procedures of PRACE, regarding the Peer Review process (documents, workflow, …) or security/storage of data policies of the Tier-0 centres involved in such pilots.

Industrial pilots were proposed by WP5 to the Board of Directors (BoD) of the PRACE AISBL and the BoD assessed, with targeted Tier-0 centres, both the legal and technical feasibility of the request. All of the pilots requesting CPU hours on PRACE have been granted through PRACE Preparatory Access calls. This was because the BoD wanted to keep the

same mechanisms for enrolling industrial pilots as for regular users and because such pilot activity was very close to the activity performed in such Preparatory Access calls.

During the project only Open R&D pilots have been possible to setup. This was mainly due to the on-going deployment of multiple Tier-0 systems with architectures which fit the needs of the pilots and the fact that pre-competitive R&D or commercial activities were almost impossible to assess due to the status of the AISBL and the Hosting Members/Sites with regard to tax and other issues.

The following table describes the six different pilots, which have been enrolled during the PRACE-1IP project:

Company Name	Country	Industrial domain	Type of company	Service assessed	Resources allocated	Duration of the pilot
CENEARO	Belgium	Aeronautics	Private / public R&D center	CPU hours with help from the center	1.3 Mh on JUGENE on up to 16 000 cores	6 months
Audionamix	France	Digital Media	SME	Advanced training	Training	One week
Vratis	Poland	CFD	Small ISV	CPU and GPU hours	200 000 CPU hours and 100 000 GPUS hours on CURIE	6 months
Thermofluids	Spain	CFD & Heat transfer	Service company	CPU and GPU hours	200 000 CPU hours and 100 000 GPUS hours on CURIE	6 months
Dompé	Italy	Life Sciences	SME + academia	CPU hours	200 000 CPU hours on CURIE	6 months
ANSYS / Tetrapack	Sweden	Food Industry	Large company + large ISV	CPU hours	Xxx CPU hours on Hermit	6 months

Table 2 - Summary of industrial pilots involved

2.4.1 *CENAERO pilot*

Regarding the first pilot with CENAERO, the first contact with this company occurred just after a face-to-face meeting with their CTO in France during a CFD workshop. CENAERO is a Belgium public/private R&D for aerodynamics which works both for academia and industry (with customers like SAFRAN). They perform R&D activity as well as studies on fluid dynamics around turbo machines or wings based on their in house CFD code called ARGO. CENAERO already used big HPC systems through the DEISA DECI calls and they had already worked closely with HPC experts from IDRIS for porting their application to IBM BlueGene/P systems.

For this pilot, CENAERO wanted to port to a PRACE BlueGene system a new version of their code implementing new CFD features. These features were to be applied to a DNS study of a low pressure blade, optimising it and scaling it out up to 16 000 cores. Contacts were established with Juelich Supercomputing Center, one of the three Tier-0 centers from GCS in Germany and after the green light from both Juelich and the PRACE BoD and the commit-

ment of CENAERO to publish the results after their grant (condition of the Open R&D access), the first Open R&D pilot started.

During the following 6 months, engineers from CENEAERO worked with the help of HPC experts from Juelich on porting, optimising collective communications and I/O on BlueGene/P and scaling out the ARGO code using real industrial test cases from Safran. The results were very good in term of performance of the code on up to 16 000 cores and K. Hillewaert, the principal investigator from CENEAERO presented the results of this pilot during the forth industrial seminar in Bologna in April 2012.

The company is very pleased about these results, and have expanded their HPC facilities with a new BULLx86 cluster for internal studies. They also plan now to use PRACE facilities again for Open R&D activities like they did during the pilot. In order to represent future industrial users, K. Hillewaert has kindly agreed to be part of the board of the PRACE Users Forum, which was setup in 2011 by PRACE-1IP WP4.

2.4.2 *Audionamix pilot*

The second pilot with Audionamix was consecutive to the engagement of this SME into a French initiative called “HPC-PME“ supported by GENCI, INRIA (the French IT R&D center) and OSEO (French public bank for innovation). Audionamix is a young SME of 20 people specialising in sound processing, more precisely for demultiplexing sound sources in real time. They are developing their own software and they wanted to be able to do it in quasi real time but by using GPUs instead of expensive solutions which are currently based on FPGAs.

The company had internal skills in C/C++ programming but no skills in GPU programming so they were enrolled as training pilot into the PRACE Autumn School organised by GENCI and CEA on “Advanced Hybrid Programming“ during October 2011. Over four days they received intensive teaching about GPU architecture and programming especially on OpenCL and HMPP. This training was very valuable for them since their main developer acquired the required skills to implement a hybrid version of the code based on OpenCL. This pilot illustrated the potential of PRACE for providing training services to industry and engaging users with emerging technologies by delivering to them standard or advanced HPC programming lectures.

2.4.3 *VRATIS pilot*

Vratis is a young and very dynamic Polish service company dedicated to developing hybrid optimised solvers (called SpeedIt and running on GPUs) for open source applications like OpenFoam. Their CEO, L. Miroslaw, gave a talk during the third PRACE industrial seminar in Stockholm and in a follow-up meeting was interested in accessing under Open R&D (performance evaluation) the PRACE facilities and especially the hybrid fraction of CURIE, the French Tier-0 system.

After some iterative discussions with the company and the BoD they applied to the preparatory access call on CURIE in March 2012 for six months access. They plan to evaluate the scalability of their SpeedIt solver library on up to 288 GPUs of CURIE and then to publish the results of this performance evaluation.

2.4.4 *Thermofluids pilot*

Thermofluids is a Spanish company specialising in the numerical simulation of computational fluid dynamics and heat transfer phenomena (CFD & HT), and its application to the simula-

tion of the behaviour or thermal systems and equipment. Thermofluids is developing an in-house parallel-unstructured CFD code for turbulent industrial problems and during this PRACE Open R&D pilot they wanted to assess the scalability on up to 8000 CPUs as well as multiple hybrid nodes.

After some iterative discussions with the company and the BoD they applied to the preparatory access call in March 2012 for six months access on the thin and hybrid partitions of CURIE.

2.4.5 *Dompé and University of Parma joint pilot*

Dompé is an Italian, company, which has already had some collaboration with CINECA one of the four PRACE hosting members. This company is developing, with the University of Parma, a drug design code called LiGen and during this Open R&D joint pilot they were interested in assessing the scalability of their code and in creating a new open database of protein crystal structures (around 2000 proteins) based on a new design methodology. This methodology is a multi-objective optimization process based on high throughput virtual screening implemented in the LiGen.

Through this joint pilot with academia, computation runs will be carried out by the University of Parma in a close collaboration with Dompé. The target system was Curie because the code was developed and already validated on similar architectures. After some iterative discussions with the company and the BoD they applied to the preparatory access call in March 2012 for six months access on the thin partitions of CURIE.

2.4.6 *ANSYS and Tetrapak join pilot*

The last pilot ran after discussions and face-to-face meetings with ANSYS, one of the biggest worldwide ISVs in scientific computing, just after the third industrial seminar in Stockholm. During the last two years, ANSYS has shown a growing interest in regard to the deployment of the PRACE RI and the process that has been started on order to engage industrial users with the infrastructure. After the seminar they proposed to set up a joint pilot with one of their key European customers, Tetrapak, one of the worldwide leaders in processing and packaging solutions for food and beverages.

This Open R&D pilot is very close to what ANSYS is doing in a more regular way in the US through the DoE INCITE program with General Motors. The goal of this pilot was to perform on Hermit, a Cray XE system at HRLS (one of the three Tier0 systems from GCS in Germany) a scalability study of Fluent, the leading CFD software application from ANSYS using industrial geometries coming from Tetrapak. After some iterative discussions with the company and the BoD they applied to the preparatory access call in March 2012 for six months access on Hermit.

2.5 Conclusion on the industrial pilots

These six Open R&D pilots performed on three different Tier-0 systems following the deployment of the PRACE Research Infrastructure, have been an opportunity for the PRACE AISBL to send a message to industry prior to the adoption of a full Open R&D offer to European industrial communities. In particular this has encouraged PRACE to adapt its operational procedures, to evaluate the adequacy of such services for industry, and to establish valuable joint partnerships between academia and industry and between ISVs and end users.

The CENAERO and the Audionamix pilots brought a real commercial advantage to these companies as they allowed them to optimise and scale out their own applications or to gain very important HPC skills for developing their business.

The others four pilots began in March 2012 but we are very hopeful that they will fruitfully develop services for ISVs and illustrate how academia and industry can jointly benefit from PRACE resources in a technology transfer context.

Since industrial users are now eligible to access PRACE Tier-0 resources for Open R&D, this pilot activity needs to be followed by PRACE at the Tier-1 layer. This layer represents the best ladder for engaging industrial users into the PRACE Research Infrastructure.

3 Definition of potential industrial targets and high value HPC services

During the PRACE-1IP project, Work Package 5 (WP5) defined the target groups of the PRACE industrial offer and studied the business models and associated high value services which could be provided to European companies interested in benefiting from the PRACE facilities.
The investigation was performed through an analysis of existing industrial offers in place in similar research infrastructures. This analysis was complemented by meetings and surveys with potential industrial users, as well as with a technological and legal assessment of each offer.

3.1 Potential PRACE industrial target groups

The following map describes what may be the different target groups of PRACE industrial offer with their motivation for using high value services:

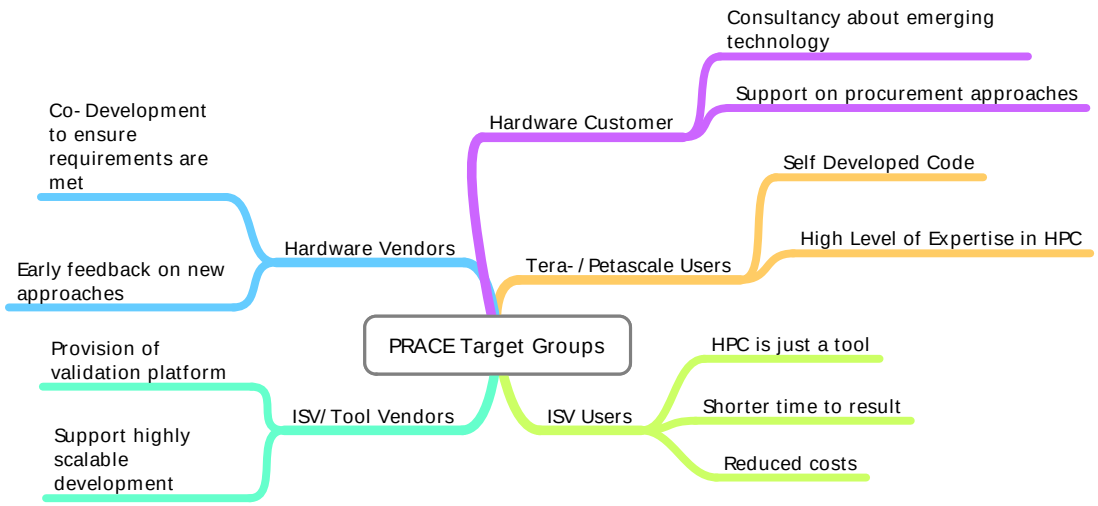

Figure 5 - Map of industrial target groups

3.1.1 *Tera/Peta scale users*

This category of users develops their own in-house codes and generally has internal HPC skills and expertise. The unique resources offered by PRACE Tier-0 systems enables new challenges in terms of problem size, complexity or time to result to be tackled. These users benefit from access to Tier-0 resources and services by accelerating their innovation process. Their costs are reduced as they can rely on an existing infrastructure and system architectures defined in collaboration between hardware vendors and hosting partners' experts (assuming that the design of a Tier-0 system architecture is a non-trivial task), instead of investing on their own large HPC resources.

The major interest of this target group is at a pure infrastructure level consisting of hardware and the corresponding system software stack. Typical collaborations may be extended in the

context of Open R&D usage via a model of academic collaboration in order to address scalability limitations or implementing new functionality.

3.1.2 *ISV User application*

This user group relies mostly on ISV (Independent Software Vendors) or Open Source codes that are not developed by the group itself. Typically, adaptation or customization of codes is performed on demand by the ISV vendor to address individual needs. Their key interest in using a very large computing system is driven by solving larger problems or solving problems more quickly – perhaps through running a series of ensembles. Their use of the PRACE Tier-0 resources is therefore mostly at the **application level** and they have little HPC expertise.

In general this user group relies on the existing capabilities of the applications and is consequently limited to Tier-2 / Tier-1 resources (due to poor scaling for example) or is aiming for a Tier-0 system in collaboration with other partners such as the ISV Vendor or in the case of Open Source applications with researchers in the application domain of the application. In this latter model the researchers, from computer science, mathematics, etc, work in a co-design methodology to target the application for the Tier-0 system level performance.

3.1.3 *ISV*

For ISVs, a major showstopper for porting an application to new hardware architectures and to develop increased scalability is the lack of access to very large computing systems under reasonable conditions. Additionally the investment in adaptation to Tier-0 level needs a certain level of committed customer demand. The interest of an ISV vendor in PRACE is on pure **infrastructure level** ideally in collaboration with an ISV application user.

This collaboration may be extended in the context of the Open R&D usage a model of academic collaboration in order to address scalability limitations or the implementation of new functionality. As there are still a large number of ISV codes that barely scale beyond a couple of compute nodes this is not an activity exclusive to Tier-0 and could be better addressed by the Tier-1 layer.

3.1.4 *Tool Vendor*

Similar to the case of the ISV, the lack of access to very large computing systems under reasonable cost conditions may be preventing investment. In contrast to the ISV vendor case, the potential customers might be the Tier-0 resource providers themselves rather than an external user. The interest of the Tool vendor is purely on the **infrastructure level** while validation of the achievements will be done using applications.

As tool support for specific system designs is to a large extent orthogonal to the system size this is not an activity exclusive to Tier-0.

3.1.5 *Hardware vendor*

Hardware vendors have interests in the validation of experimental or innovative approaches in system design or software solutions in order to assess their applicability for future Tier-0 resources. The major interest in an industrial offer from PRACE (beside procurement of production systems) is the joint validation and development of future system architectures of Tier-0 systems. The interest is not limited to Tier-0 resources.

In the field of the upcoming PRACE Third Implementation Phase (PRACE-3IP), a Pre Competitive Procurement (PCP) for HPC will be evaluated by PRACE. This new mechanism concerns the Research and Development (R&D) phase before commercialization of a new product and R&D can cover activities such as solution exploration and design, prototyping, up to the original development of a limited volume of first products or services in the form of a test series.

3.1.6 *Hardware customer*

While the access to Tier-0 resources might be applicable for a large number of users, access by some commercial companies may not be an option due to constraints such as security or privacy. Consequently the interest of access to Tier-0 resources is not for long term usage of the resource but for access to a variety of different systems for validating and analysing different types of hardware – perhaps as part of a procurement process. These may be either Tier-0 production systems or experimental prototypes operated by a PRACE partner.

3.2 High value HPC services

Based on the gathering of industrial users' needs presented in D5.1.2, the PRACE projects have developed an initial list of potential high value services. They will complement a rough PaaS (or even IaaS for infrastructure level) access to Tier-0 resources and will represent a major differentiator which will foster the innovation process by bridging the gap between academia and industry.

This need of high value services from PRACE has been highlighted multiple times during the different industrial seminars. PRACE needs to be an enabler of technology transfer between academia and industry. These services are organised into the following four sections, which will allow industrial users to benefit from the expertise accumulated by the PRACE partners over the past three decades.

The description of the services below does not consider costs but focuses only on the content of the service. It is obvious that some services cannot be delivered for free and that appropriate funding for staff and facilities is required to deliver the services. In particular the Software Services toolbox element is typically delivered as consultancy service and represents a major element of existing industrial offers from the PRACE partners.

3.2.1 *Access to HPC resources*

Accessing various and massive HPC resources from PRACE at the Tier-0 and the Tier-1 levels has been one of the main offerings that captures the interest of companies during PRACE industrial seminars.

These events have been attended by companies (large companies and SMEs) developing their own codes and able to run at the Tier-0 level with capability simulations on more than 2000 cores. A lot of companies have raised the need for PRACE to not only consider capability

simulations but also a mix of capacity and capability for addressing uncertainties quantification (based on parametric studies) or coupled multi-physics simulations.

This need is very close to the needs expressed by academia for running big ensemble simulations and operational models from PRACE centres have evolved in that direction. As an example on CURIE (from GENCI@CEA) and SuperMUC (from GCS@LRZ) it will be possible to run per user between up to 20 to 50 simulations of 512 cores at the same time (depending on the load on the machine).

On the other side SMEs and even large companies relying on ISV codes or with little HPC experience are not eligible directly at the Tier-0 level because they tend to rely on simulations on less than 128 cores.

For this category of companies, which is by far the biggest one, Tier-1 in PRACE may play a central role for promoting, attracting, engaging and providing high value services to such users.

3.2.2 *Training and information services*

PRACE may propose a dedicated industrial section for the PRACE RI website and may welcome the participation of ISV and Open Source developers to the online training portal and to the PRACE Advanced Training Centres (PATC). This participation may allow PRACE to deliver dedicated content targeted to industrial users who are usually most interested in best practices for using off-the-shelf software in the most efficient way.

Multiple ISVs have expressed a wish to be part of the PRACE training programme during the seminars or during face-to-face meetings. As pilots, two ISVs Allinea (a UK based company delivering parallel debugging and profiling tools) and CAPS Enterprise (a French company providing tools for heterogeneous many-core programming) have been enrolled as lecturers during PRACE seasonal schools. They accepted the rules of coming to the seasonal schools for giving technical lectures in their domain without any commercial speech and no funding for their participation. Such examples may be augmented with other ISVs interested in other science fields such as CFD, structural mechanics, acoustics, code coupling, …

With regard to information services, this issue may be viewed as simplistic but it is very important for a lot of European companies (from large companies to SMEs) who recognize that they lack information about the potential of HPC in order to increase their competitiveness They need information on: how to be trained, how to access to HPC resources, which methodologies/tools to use, what services are provided, … and more globally what is the European HPC ecosystem.

3.2.3 *Software services*

Industrial users may benefit from the existing PRACE calls for proposals like the “PRACE preparatory access”. This allows organisations to be granted access on PRACE Tier-0 systems and to be helped by PRACE experts for porting and performing software optimisations on their code prior to submitting proposals to “regular calls”.

At the earliest stage of the development, industrial users may work in a co-design mode on developing their software in integrated teams. Theses teams may be formed by industrial experts working jointly with experts from academia (in numerical simulation, meshing, software development, …), HPC experts and eventually HPC vendors. Such work fosters technology transfer and tackles the new challenges arising from the ever greater complexity of HPC hardware and software architectures.

Industrial users may also be involved in PRACE activities on code enabling for community Open Source codes for industry similar to what is done in PRACE-2IP WP9 with OpenFoam (one of the leading CFD open source code) for example, by providing skills or industrial geometries for driving and validating the code enabling activity.

3.2.4 *Technology watch*

Interested industrial users may benefit from the R&D activity and the technology watch activities provided by the PRACE partners.

This may lead to their participation in the hardware and software prototyping activities in collaboration with the PRACE partners in order to anticipate emerging technologies or to access best practices guides and white papers about key issues like Green Computing, future programming languages, best practices for HPC procurements, etcetera.

3.2.5 *Co-Development*

In addition to technology watch activities, PRACE partners can establish a partnership with vendors in hardware and software for the development of advanced products and services based on joint research and development activities.

As of May 2012, the creation of a European Technology Platform (ETP) in the area of HPC is still in progress with a potential incorporation date in either May or June 2012. The ETP's objective is to prepare a European research agenda that will help the European HPC solution provider ecosystem remain competitive at a global level. In this respect, PRACE will focus on the provision of a pan-European HPC research infrastructure, while the ETP will be responsible for defining Europe's research priorities in the field of HPC technology development. As PRACE advances its industrial services model (e.g. through the Open R&D Access Model), the ETP should take advantage of such programmes in order to identify potential end-user organisations that might be interested in working within the framework of the ETP. One of the key requirements for joining the organisation is having an HPC research portfolio in Europe and it is expected that some of PRACE's industrial partners will both fulfil this criterion and be interested in developing a European Strategic Research Agenda in the area of HPC. In this regard, PRACE should develop a mechanism for sharing such information with the ETP.

5 Assessment of potential business models for industry

Based on the gathering of industrial users needs presented in D5.1.2 and the assessment of existing business models presented in D5.2.1, three models have been selected in order to fit the needs of potential industrials users. These are:

5.1 Open R&D

In this business model industrial users may only use the facilities and services provided by the infrastructure for basic research and development purposes. The condition associated with this free access is for the industrial user to publish all results obtained at the end of the grant period. Industrial users may apply as principal investigator on their own or in some cases may be paired with an academic partner in order to foster technology transfer between academia and industry.
This Open R&D business model is already used worldwide by some other HPC research infrastructures such as the DoE INCITE program in USA or in Europe by GENCI (France).

Based on the outcome of the industrial pilots and the strong interest from industrial users contacted during seminars and face-to-face meetings, a decision of the PRACE AISBL Council was taken during summer 2011 to ask WP5 to finalise a PRACE Open R&D offer.

In order to assess the possibility that the PRACE AISBL support such a business model in terms of legal and financial issues, the status of the AISBL and the terms of Open R&D have been investigated in collaboration with WP2 with the support of Bird&Bird (who provide legal support to PRACE).

In addition, this assessment has been performed amongst the four PRACE Hosting Members (Germany, France, Italy and Spain) and their corresponding sites (JSC, HLRS and LRZ for GSC in Germany, CEA for GENCI in France, CINECA in Italy and BSC in Spain).

WP5 also worked in a close relationship with the BoD (Board of Directors) and SSC (Scientific Steering Committee), two internal bodies of the PRACE AISBL, in order to define how Open R&D may be proposed in terms of organisational issues.

An initial decision was taken not to have a separate call for proposals for industry but to merge industrial users with existing users from academia into existing PRACE calls for proposals based only on the criteria of scientific excellence. Industrial users may become eligible to PRACE Preparatory Access calls (from two to six months allocation for porting and optimising their application in order to be prepared for a regular call) and PRACE Regular Calls (one year allocation for performing scientific simulations).

Industrial users may apply to PRACE calls for proposals alone as a single PI or paired with academia in a joint public/private collaborative project. This last point has been highlighted during the PRACE industrial seminars where attendees saw PRACE as an enabler for engaging industrial users in HPC and fostering technology transfer between academia and industry.

Since PRACE calls are peer reviewed based on the single and unique criteria of scientific excellence by a PRACE external panel of scientific experts, the SSC assessed the need to modify existing panels and committees in order to take into account the arrival of industrial oriented projects.

The result has been that industrial users benefit from the same SLA (Service Level Agreements) as academic users in term of levels of support, confidentiality of data, storage policies, etc.

In order to more efficiently support specific usage from industry such as uncertainty studies or multi-physics simulations, capacity jobs will be supported by the Tiers-0 centres provided that individual simulations use at least 512 cores.

This usage model was also requested by scientific communities in order to support so called ensemble (multi-physics) simulations or parametric studies.

Based on these results WP5 developed an Open R&D business model which has been proposed to and voted on at the 6th Council meeting of PRACE AISBL, in Copenhagen, January 24th 2012:

"The industrial usage of Open R&D should be based on the following principles:

- *Access to PRACE AISBL Tier-0 systems and services opened to industry users exclusively for R&D projects*
- *Scientific excellence will be the unique selection criterion*
- *Scientific results must be published at the end of the allocation period*
- *Access for Industry will be free of charge but conditional*

The organisational and technical aspects for Open R&D includes:

- *Industry users included into current PRACE calls (Preparatory Access and Regular calls)*
- *Industry may bring their own software licenses if needed*
- *No need to change SSC or AC composition and management except to add industry experts into review panels*
- *Industry and academic have same service levels*

Based on the assessment of the Open R&D business model, it is proposed to the Council to extend the access of PRACE calls to industrial users with the following rules for the next two years:

- *Companies having HQ or substantial R&D activity in Europe are eligible*
- *Industrial users may apply as single investigator, or paired with an academic partner*
- *Each access as single contributor will be limited to a maximum of 5% of the total computing resources of a single PRACE system, subject to the approval of the boundaries imposed by EC.*
- *Prior to being granted, companies will commit on publishing the results at the end of the grant period*
- *Access will be free of charge"*

As a result, it was announced during the 4th PRACE Industrial Seminar in Bologna (April 16-17th 2012) that PRACE now allows, starting from the 5th Regular Call and June's Preparatory Access deadline, industrial users to access its resources in order to conduct open research.

Within the PRACE peer-review process, the same rules apply as for academic users: scientific excellence is the only selection criteria (in order to avoid distorting the market). The results of projects that involve industrial organisations must remain open – i.e. they must be made available publicly on project completion. The limitation of 5% of the total computing resources of a single PRACE system, for industrial single contributors, both complies with legal requirements and gives incentives for industrial users to be paired with an academic partner and therefore foster technology transfer.

This decision of the PRACE AISBL for the Tier-0 centres has been partially extended to the Tier-1 centres of PRACE participating in the DECI calls. There is also an ongoing proposal which is being discussed between WP4 and WP5 for extending the full Tier-0 Open R&D

model to Tier-1. This proposal is included into the D4.3.2 “Cross-national programme for Tier-1 access pilots ».

This support of Open R&D is a natural step as Tier-1 partners represent a key ladder for engaging industrial users, especially SMEs. Such companies have very often less knowledge of HPC and fewer scalable codes than academia, a lower maturity for using a high number of computing resources or they use costly ISV codes which are too expensive for use on core counts.

This also represents a desire of the infrastructure to expose to potential users (from academia to industry) a coherent and scalable vision of the business models and services across the pyramid of European HPC resources starting from Tier-2 centres (regional centres), to Tier-1 centres (national centres) and finally Tier-0 centres (European centres).

In order to efficiently engage companies (especially SMEs) who have little expertise and knowledge about the potential of HPC to increase their competitiveness, new complementary services need to be developed around the Open R&D business model. These upstream services will take benefit from the expertise of PRACE staff as well as researchers from scientific communities in order to inform, train, understand what the company is doing, co-design an industrial project using HPC and demonstrate it on PRACE resources.

This “evangelisation” program, which focuses on the benefit of HPC with regard to driving innovation and competitiveness, can be seen as an integrated set of complementary services. As shown during the fourth industrial seminar, similar initiatives exist at the national level in Scotland (the Supercomputing Scotland programme by EPCC) and in France (the HPC for SMEs Programme by GENCI, INRIA and OSEO). PRACE plans to evaluate the possibility of scaling out such programs at the European level in the scope of the PRACE-3IP (Third Implementation Phase) project foreseen starting in July 2012.

5.2 Pre-competitive R&D

In addition to Open R&D another model exists for industrial users in order to access the facilities and the services of the infrastructure for performing software/product development in a pre-competitive stage and in a less restrictive way.

Compared to Open R&D the condition to publish the results immediately is weakened. In some cases a delay of up to three years is given to the industrial user before publishing the results in order to provide time to ship the product and to take the benefits of its innovation. At an operational level, there are three ways for implementing such a business model:

1. By using the multi year access currently evaluated by PRACE RI into Open R&D calls, this will allow the company to defer the full publication of the results for 2 to 3 years. In order to be eligible, the company will need to prove that its R&D program span for the duration of the multi year access call and commit to communicate internally to the PRACE Peer Review panels of experts during yearly reviews about intermediate results.
2. By using a regular one year grant but asking for deferring the publication of the results. These conditions are not compatible with the Open R&D business models or more generally with conditions of PRACE existing calls since they create a distortion between academia users obliged to publish results at the end of the grant period and some industrial users who are allowed to postpone the delivery of the results or to not even publish them. If selected by the PRACE AISBL this possibility could be implemented into a new kind of call based on scientific excellence on a limited set of resources with an associated fee to defer the publication of the results.

3. Finally, a last possibility offered by Pre Competitive R&D access is to perform R&D activity without publishing the results at the end of the grant period. This business model is very often associated with commercial activity since the access and the services are charged at the market price. On some cases the cost of such services may be subsidized if the developments represent a potential and direct interest for the infrastructure or if, for example, they have a societal impact. This model requires very careful observation of the use of the resources and clear processes to control the compliance with the delayed publication and potential measures to enforce them.

All organisational constraints of Open R&D apply similarly to pre-competitive R&D with the additional challenge that the boundaries between Open R&D, Pre-Competitive R&D and Production use are very difficult to define unambiguously. The major additional problem is that all accounting, billing and monitoring activities including the acceptance of Service Level Agreements, penalties on the provider side in case of problems of data security, of confidentiality of results, or of unavailability of resources, do apply similarly to the full commercial production offer. In particular the delayed publication procedure might, in case of a violation by the customer, require a treatment of the activity that is, retrospectively, similar to the one for production activities.

As in the Open R&D business model, the adequacy, in terms of legal and financial issues, of Pre Competitive R&D with PRACE AISBL statutes as well as with the four PRACE Hosting members and the corresponding hosting sites have been investigated with the help of WP2. Results showed that such access may be possible in the framework of the Open R&D multi-year access (as mentioned in scenario 1 above) but more complicated for the 2 others scenarios due to current PRACE statuses, with a high risk in terms of market distortion, dominant position for PRACE, VAT issues, etc.

5.3 Production, consultancy and commercial activities

At this stage the need of the industrial user is to use the infrastructure in production mode for commercial activities. This needs to be charged at the market rate. Usage may be regular, with strong associated levels of service in terms of security, confidentiality, and availability of the systems. Conversely, it may be irregular, with constraints for provisioning of on-demand HPC facilities following agreement with the end user.

Often the specific environment demands are defined in a joint certification process between OEMs and supporting suppliers limiting significantly the options in terms of OS, middleware, hardware and applications that can be provided.

The provision of such a service on public funded resources require an appropriate administrative management of the provision in terms of contracts, taxes etc. in order to comply with the regulations to provide such services on publicly funded resources.

When systems are eligible for supporting commercial activities (there may be some limitations due to the statuses of HPC centers, the access to academic networks, the non commercial use of HPC resources required by the vendors, …) typical approaches are to estimate and declare a certain fraction of the system e.g. 15% to be used commercially and monitor and review the real use and adapt accordingly. However for such eligible systems there is no actual limitation about the fraction that is dedicated to commercial production use, which might reach up to 100% of the system.

Orthogonal to the provision of HPC services and solutions based on a hosted Tier-0 system, a contribution in terms of knowledge-transfer, consultancy and training can be realised aimed at shortening the time until an innovation in the academic space is adopted by industry.

For this kind of offer all constraints of a full commercial offer apply in common with any commercial provider of computing and HPC services. As mentioned above this applies to all organisational activities such as resource usage monitoring, billing, contract management, tax declaration and the acceptance of Service Level Agreements on the provider side that typically contain penalties in case they are not respected. The risk taken by the provider must be compensated by corresponding income and is realised using market prices for the offering.

Since usually only a part of the system is made available for commercial use, all costs that occur for operating the system such as investment costs, operations costs (power, cooling, staff), user support costs or specific costs like:

- Additional facility security activities (e.g. camera surveillance, more security staff, specific data isolation, system access with strong data encryption, …)
- Additional insurances
- Administrative and Legal Support
- Licenses for tools and applications for commercial use
- Specific hardware and services (e.g. specific filesystems)
- Tax declarations, …

need to be considered keeping in mind that only a fraction is used for commercial activities. Some costs may be shared (power), some may not (specific security or strongest SLA), but in any case, these costs need to be clearly documented and subject to a clear policy.

Such commercial activities are already proposed as shown in D5.2.1 by HPC research infrastructures in Europe in PaaS or SaaS mode (CINECA in Italy, EPCC in the UK, …) or in the USA with NCSA and the Private Sector Program (PSP).

In common with the Open R&D business model, the fit of commercial activities with the PRACE AISBL statutes in terms of legal and financial issues as well as with the four PRACE Hosting members and corresponding hosting sites have been investigated with the help of WP2.

Results showed that with current PRACE statutes such a business model would be very difficult and risky to implement in terms of market distortion, dominant position for PRACE, VAT issues, etc.

HPC access for commercial activities is also already provided not only by research infrastructures but also by commercial companies such as well established HPC vendors (IBM, SGI, HP, BULL and others) but also by Cloud Services providers who want to support HPC services. The following two examples come from the Amazon Web Services and the BULL Extreme Factory offers for Cloud Services for industry:

5.3.1 *Amazon offer*

The most notable full commercial HPC offering is Amazon Web Services[1] (AWS) which started in 2006. Amazon markets a full list of Cloud services providing access to remote on demand IT services such as resource brokers, authentication, monitoring, online billing, storage, search, training, database, email, etc. The most central and well known of these services are Amazon EC2 (Elastic Compute Cloud) and Amazon S3 (Simple Storage Service).

[1] http://aws.amazon.com/

Amazon Elastic Compute Cloud (EC2) is a central part of Amazon's cloud computing platform, Amazon Web Services. EC2 allows users to rent virtual computers on which they can run their own computer applications. EC2 allows scalable deployment of applications by providing a Web service through which a user can boot an Amazon Machine Image to create a virtual machine, which Amazon calls an "instance", containing any software desired. A user can create, launch, and terminate server instances as needed, paying by the hour for active servers, hence the term "elastic". EC2 provides users with control over the geographical location of instances that allows latency optimization and high levels of redundancy.

Amazon S3 (Simple Storage Service) is an online storage web service offered by Amazon Web Services. Amazon launched S3, its first publicly available web service, in the United States in March 2006[2] and in Europe in November 2007. At its inception, Amazon charged end users US$ 0.15 per gigabyte-month, with additional charges for bandwidth used in sending and receiving data, and a per-request (get or put) charge. As of November 1, 2008, pricing moved to tiers where end users storing more than 50 terabytes receive discounted pricing. Amazon claims that S3 uses the same scalable storage infrastructure that Amazon uses to run its own global e-commerce network.

In term of costs, Amazon AWS offers three types of cost model depending on the type of instances used by the remote customer:

- On-Demand Instances – On-Demand Instances let the user pay for compute capacity by the hour with no long-term commitments. This frees the user from the costs and complexities of planning, purchasing, and maintaining hardware and transforms what are commonly large fixed costs into much smaller variable costs. On-Demand Instances also remove the need to buy “safety net” capacity to handle periodic traffic spikes.
- Reserved Instances – Reserved Instances give the user the option to make a low, one-time payment for each instance the user wants to reserve and in turn receive a significant discount on the hourly charge for that instance. There are three Reserved Instance types (Light, Medium, and Heavy Utilization Reserved Instances) that enable the user to balance the amount charged upfront with the effective hourly price.
- Spot Instances – Spot Instances allow customers to bid on unused Amazon EC2 capacity and run those instances for as long as their bid exceeds the current Spot Price. The Spot Price changes periodically based on supply and demand, and customers whose bids meet or exceed it gain access to the available Spot Instances. If you have flexibility in when your applications can run, Spot Instances can significantly lower your Amazon EC2 costs.

At the beginning such AWS services were oriented to the general IT industry and the web services but for the past two years Amazon has offered Cloud Services to the HPC market.

HPC on Amazon EC2 is enabled by the Cluster family of instance types. Cluster Compute and Cluster GPU instances can be used just like any other Amazon EC2 instance but also offer the following features optimized for HPC applications:

- Cluster instances can be launched within a *Placement Group*. All instances launched within a Placement Group have low latency, full bisection 10 Gbps bandwidth between instances. Like many other Amazon EC2 resources, *Placement Groups* are dynamic and can be resized if required. Also connections of multiple *Placement Groups* are possible in order to create very large clusters for massively parallel processing.

- Cluster Compute and Cluster GPU instances specify the underlying processor architecture in their definition to enable developers to tune their applications by compiling for that specific processor architecture in order to achieve optimal performance.
- Cluster GPU instances allow customers to take advantage of the parallel performance of NVidia Tesla GPUs using the CUDA and OpenCL programming models for GPGPU computing.

During the Supercomputing conference in November 2011 in Seattle, Amazon announced a HPC system of 17 024 cores for a peak power of 354 TFlops. The system was ranked #42 in the November 2011 edition of the top500 with a sustained performance of 240 TFlops.

This system is the flagship system of the Amazon HPC offer but other small systems based on GPU resources for example are also available in US.

Amazon has started to work with some ISVs in the biomedical market or in finance in order to sell solution based offers. However for the moment, due to the lack of high performance interconnect between the compute nodes (only based on Gb Ethernet and not Infiniband), the HPC offer is reserved for embarrassingly parallel applications or minimally coupled simulations. The price list (see http://aws.amazon.com/ec2/pricing/) is vast and dependant of the location of the resources (US, Asia, South America or Europe), the type of instance (on demand or reserved), the duration of the reservation for reserved instances and the kind of hardware resources needed (a single low performance core is cheaper than a full HPC configuration with GPUs) and the kind of OS (Windows or Linux).

The following hardware and software descriptions of the HPC instances are proposed by AWS:

- Cluster Compute Quadruple Extra Large Instance
  - 23 GB of memory
  - 33.5 EC2 Compute Units (2 x Intel Xeon X5570, quad-core “Nehalem” architecture)
    1690 GB of instance storage,
  - I/O Performance: Very High (10 Gigabit Ethernet)
  - API name: cc1.4xlarge
- Cluster Compute Eight Extra Large Instance
  - 60.5 GB of memory
  - 88 EC2 Compute Units (2 x Intel Xeon E5-2670, eight-core "Sandy Bridge" architecture)
  - 3370 GB of instance storage
  - I/O Performance: Very High (10 Gigabit Ethernet)
  - API name: cc2.8xlarge
- Cluster GPU Quadruple Extra Large Instance
  - 22 GB of memory
  - 33.5 EC2 Compute Units (2 x Intel Xeon X5570, quad-core “Nehalem” architecture)
  - 2 x NVIDIA Tesla “Fermi” M2050 GPUs
  - 1690 GB of instance storage
  - I/O Performance: Very High (10 Gigabit Ethernet)
  - API name: cg1.4xlarge

As an example in US (Virginia) the prices for HPC resources are the following (in May 2012):

- For On demand instances on Linux:

  - $1.30 per hour for Quadruple Extra Large and $2.40 per hour for Eight Extra Large
  - $2.10 per hour for Quadruple Extra Large on GPU resources
- For reserved instances on Linux with an heavy usage of resources:
  - For Quadruple Extra Large: An upfront cost of $4,060 and $0.297 per Hour for one year commitment and an upfront cost of $6,300 and $0.297 per Hour for one year commitment
  - For Eight Extra Large: An upfront cost of $5,000 and $0.361 per Hour for one year commitment and an upfront cost of $7,630 and $0.361 per Hour for one year commitment
  - Quadruple Extra Large on GPUs: An upfront cost of $6,830 and $0.494 per Hour for one year commitment and an upfront cost of $10,490 and $0.494 per Hour for one year commitment

Data transfers are also charged from $0.120 per GB (for up to 10 TB/month) to $0.0 per GB (for up to 350 TB / month).

5.3.2 *Bull offer*

At the SC'10 supercomputing conference in New Orleans, USA, Bull unveiled its Cloud solution for HPC, named « extreme factory ». It is a fully integrated pay-as-you-go computer simulation offering. Bull has a wide range of cloud solutions for all types of needs, but this one is dedicated to HPC.

It is worth noting that many ISVs (a player usually missing in these types of offers) including ESI Group, CD-Adapco, Exa and The Bakery have worked with Bull to make their applications available on the platform. “Extreme factory” - which includes a secure Web portal and both hardware and software resources – was first available in January 2011.

As with the Amazon offer, three usage models are available:

- XF Dedicated: Resources dedicated to users with a long-term need (six months or more) for compute resources; dedicated HW resources (compute and service nodes, storage) are allocated to ensure guaranteed stability and security. Customization can easily be added to this model. Optional VPNs can be either operated by the customer’s operator or by Bull.
- XF Reserved: Guaranteed resources upon reservation - lots of users have peak loads requiring resources, and compute time is allocated in periods of one or more weeks. This is achieved by dedicating virtual login nodes and provisioning the associated physical compute nodes for the duration requested.
- XF Shared: Resources are mutualized and allocated on a ‘First-In-First-Out’ basis. This is the model which is the closest to the traditional commercial ‘cloud’ model. Users access extreme factory whenever they want to, whenever they need to and are billed for the actual time used.

As for the type of resources that one can use, the list as of May 2012 is:

- Two types of compute nodes are available in the extreme factory supercomputer. They are regularly updated in order to accommodate state-of-art technology: bullx B500 and B505 blades have highly optimized HPC dual Intel® Xeon® 5600 processors with 24 Gigabytes of memory, and bullx B505 have additional dual NvidiaTM M2070 GPUs.

- A few large memory nodes based on bullx R423-E2 or bullx S6030 are also available for specific operations like pre-processing data models which require large memory configurations (512 gigabytes and more).
- Service nodes are either physical bullx R423-E2 nodes with good I/O and memory capabilities or virtual nodes for login and security isolation and R425-E2 nodes for visualization.
- All nodes run a standard Linux® environment based in RHEL 5 or 6 compatible kernels as well as Bull's own cluster management software. Windows® nodes can also be deployed on an individual project basis.
- There are two types of storage: a high-performance parallel Panasas storage cluster for scratch and a dedicated high capacity NetApp system for long term storage.
- Standard communication needs are addressed by 100Mb/s to 1Gb/s bandwidth secured lines which are accessible via the Internet with optional VPN. For customers who have higher bandwidth requirements, Bull PI (a Bull subsidiary) proposes the rapid installation of adequate point-to-point links at speeds up to 10Gbits/s.

The initial price list included:

- Between €0.20 per core, per hour and €0.30 per core, per hour depending on the number of blade servers being used and the overall usage time
- Between €0.15 per GB, per month and €0.30 per GB, per month depending on the number of storage systems involved and how long they are used for
- Consulting services: €1,300 a day
- SaaS services: varies according to publishers' tariffs

6 Conclusions

During the PRACE-1IP project, WP5 finalised the gathering of the needs and expectations of industrial users by organising two very successful seminars in Stockholm and Bologna. Such events, their follow-up as well as an extensive survey showed a strong interest from industry (from large scale companies to SMEs and ISVs) in benefiting from HPC in order to increase their competitiveness and boost the effectiveness of their value chain. PRACE is now known by industry and seen as an enabler in fostering HPC adoption in Europe and facilitating technology transfer between academia and industry.

This is one of the reasons why WP5 had focused on investigating Open R&D services and providing feedback on all relevant issues to the PRACE AISBL Council in order to decide to support this business model for industry. Similarly to what is already being done in the US within the DoE Incite program, PRACE Open R&D calls will allow European companies (or foreign companies having a significant R&D activity in Europe) to benefit from the large-scale HPC resources and services offered by PRACE through a peer-reviewed call application process with scientific excellence as the main criterion.

In order to increase the footprint of potential industry users, such calls, currently available on the Tier-0 facilities only, may be extended to PRACE Tier-1 systems, as this represents a perfect scale for engaging industrial users who are adept at simulating on smaller systems or using tools with low scalability.

Finally, another way of engaging industrial users, especially SMEs, will be to set up a new programme based on an integrated set of services for informing, training, understanding the needs of the company, co-designing an industrial project requiring HPC and demonstrating it on PRACE systems. As SMEs represent a very important component of industrial activity in Europe, a wide “evangelisation” program will attract new companies (i.e. the companies that are not aware of the potential of HPC) to the HPC technology, by complementing these services with the expertise of PRACE experts and also scientists from public research organisations.

Based on the existing examples in Scotland (the Supercomputing Scotland programme by EPCC) and in France (the HPC for SMEs programme by GENCI, INRIA and OSEO), this new programme will be investigated in the third implementation project (PRACE-3IP) supported by the European Commission.