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

D8.4
Final Report on Petascale Systems and Centre Assessment

Final

Version: 1.1
Author(s): Guillermo Aguirre, BSC
François Robin, Jean-Philippe Nominé, CEA
Marco Sbrighi, CINECA
Date: 25.06.2012

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: D8.4	
	Deliverable Nature: Report	
	Deliverable Level: PUBLIC	**Contractual Date of Delivery:** 30 / 06 / 2012
		Actual Date of Delivery: 30 / 06 / 2012
	EC Project Officer: Thomas Reibe	

Document Control Sheet

Document	**Title: Final Report on Petascale Systems and Centre Assessment**	
	ID: D8.4	
	Version: 1.1	**Status:**Final
	Available at: http://www.prace-project.eu	
	Software Tool: Microsoft Word 2010	
	File(s): D8.4.docx	
Authorship	**Written by:**	Marco Sbrighi - CINECA Jean-Philippe Nominé, François Robin -CEA G. Aguirre - BSC
	Contributors:	I. Liabotis, G. Karagiannopoulos -GRNET L. Gilly - CSCS G. Svensson - SNIC
	Reviewed by:	J. Wolfrat, SARA D. Erwin, FZJ (PMO)
	Approved by:	MB/TB

Document Status Sheet

Version	Date	Status	Comments
0.1	May 18, 2012	Draft	Outline and global structure + first contributions on Infrastructure topics (Ladina and JP)
0.2	May 23, 2012	Draft	Merged sections on Sources/Tools (Ioannis) + Procurements (Marco) + augmented references list
0.3	May 28, 2012	Draft	Merged contributions on Market Survey and hw/sw correlation (Chapters 2 and 3 by Guillermo)
0.4	June 8, 2012	Draft	Complements Chapters 2/3 Draft Exec Summary, introduction and conclusion
1.0	June 11, 2012	Final	Final (temporary) version for PRACE internal review
1.1	June 22, 2012	Final	Corrections after review, version for final approval

Document Keywords

Keywords:	PRACE, HPC, Research Infrastructure HPC market, commissioning, procurement, data centre

Disclaimer

This deliverable has been prepared by Work Package 8 of the Project in accordance with the Consortium Agreement and the Grant Agreement n° RI-261557. It solely reflects the opinion of the parties to such agreements on a collective basis in the context of the Project and to the extent foreseen in such agreements. Please note that even though all participants to the Project are members of PRACE AISBL, this deliverable has not been approved by the Council of PRACE AISBL and therefore does not emanate from it nor should it be considered to reflect PRACE AISBL's individual opinion.

Copyright notices

Table of Contents

List of Figures

List of Tables

References and Applicable Documents

[1] http://www.netvibes.com/
[2] http://www.google.com/cse/
[3] http://www.netvibes.com/hpc-market-watch#HPC_Market_Watch
[4] http://www.google.com/cse/home?cx=000869963287342404080:2llwvbxdrbo
[5] Earl Joseph, Steve Conway and Jie Wu, "IDC's Top 10 HPC Market Predictions for 2010", February 2010 (http://www.hpcuserforum.com/EU/downloads/IDC_HPCTop10Predictions2010.pdf)
[6] "Trends and Opportunities in High Performance Computing (HPC)" (http://agendabuilder.gartner.com/lsce6/WebPages/SessionDetail.aspx?EventSessionId=783)
[7] Earl Joseph, Steve Conway, Jie Wu, Lloyd Cohen and Charlie Hayes, "IDC HPC Market Update", June 2010(http://www.earljoseph.info/IDC_HPC_market_overview_FINAL.pdf)
[8] Christopher Willard, Ph.D., Addison Snell, Sue GouwsKorn, CFA, Laura Segervall, "Worldwide High Performance Computing (HPC) 2010 Total Market Model and 2011-15 Forecast: Overview", April 2011 (http://www.intersect360.com/industry/reports.php?id=49)
[9] "HPC Industry Dynamics",(http://www.intersect360.com/industry/)
[10] HPC Market Update by Intersect360 Research at SC11 (presentation by Addison Snell)
[11] IDC HPC Market Update, 42nd HPC User Forum Meeting, September 2011
[12] IDC's Top 10 HPC Market Predictions for 2012, February 21, 2012 - http://event.on24.com/event/39/34/56/rt/1/documents/slidepdf/wc20120221.pdf
[13] http://ec.europa.eu/invest-in-research/pdf/download_en/risk_management.pdf
[14] http://ec.europa.eu/information_society/research/priv_invest/pcp/index_en.htm
[15] http://cordis.europa.eu/fp7/ict/pcp/home_en.html.
[16] http://www.ertico.com/assets/Activities/P3ITS/P3ITS-D2.1-Analysis-of-public-Pre-Commercial-Procurementv1.8.pdf
[17] http://www.pcp2011.eszakalfold.hu/
[18] http://www.ertico.com/pre-commercial-public-procurement-for-its-innovation-and-deployment
[19] http://www.ertico.com/assets/Activities/P3ITS/P3ITS-D2.1-Analysis-of-public-Pre-Commercial-Procurementv1.8.pdf
[20] http://www.wto.org/english/docs_e/legal_e/gpr-94_e.pdf
[21] PRACE Preparatory Phase Deliverable 7.1.3 "Final assessment of Petaflop/s systems to be installed in 2009/2010", June 2009 - Jean-Philippe Nominé
[22] PRACE Preparatory Phase Deliverable 7.4.1,"Initial Risk Register", December 2008 - Horst-Dieter Steinhöfer
[23] PRACE Preparatory Phase Deliverable 7.4.2, "Final Risk Register", December 2009 - Horst-Dieter Steinhöfer
[24] PRACE Preparatory Phase Deliverable 7.6.1, "Procurement strategy", December 2008 - R. J. Blake
[25] PRACE Preparatory Phase Deliverable 2.3.1, "Document on Procurement strategy", April 2008 - Stefan Wesner
[26] PRACE-1IP Confidential Deliverable 8.1, "Preliminary report on Petascale systems and centre assessment", December 2010 – F. Robin
[27] PRACE-1IP Confidential Deliverable 8.2, "Updated report on Petascale systems and centre assessment", June 2011 – G. Aguirre
[28] PRACE-1IP Confidential Deliverable 8.3, "Consolidated report on Petascale systems and centre assessment", December 2011 - N. Meyer, R. Januszewski, D. Kaliszan

[29] Selecting a Location for an HPC Data Centre, PRACE-1IP WP8 White paper - Ladina Gilly (www.prace-ri.eu/whitepapers)

[30] Cooling – making efficient choices, PRACE-1IP WP8 White paper - Radosław Januszewski et al. (www.prace-ri.eu/whitepapers)

[31] Electricity in HPC Centres, PRACE-1IP WP8 White paper - Marcin Pospieszny (www.prace-ri.eu/whitepapers)

[32] Redundancy and reliability for an HPC Centre, PRACE-1IP WP8 White paper - Erhan Yilmaz (to be published, www.prace-ri.eu/whitepapers)

[33] Security in an HPC Centre, PRACE-1IP WP8 White paper (to be published, www.prace-ri.eu/whitepapers)

[34] HPC Systems Procurement Best Practice, PRACE-1IP WP8 White paper – Richard Blake et al. (www.prace-ri.eu/whitepapers)

[35] PRACE RI Web page for White Papers www.prace-ri.eu/whitepapers

[36] http://www.hpcresearch.nl/euroben/reports/

List of Acronyms and Abbreviations

AC	Alternating Current
ACML	AMD Core Math Library
AMD	Advanced Micro Devices
AVX	Advanced Vector Extensions
BSC	Barcelona Supercomputing Center (Spain)
ATI	Array Technologies Incorporated (AMD)
BAdW	Bayerischen Akademie der Wissenschaften (Germany)
CEA	Commissariat à l’Energie Atomique et aux Energies Alternatives (represented in PRACE by GENCI, France)
CEA/DAM	CEA/Direction des Applications Militaires
CINECA	ConsorzioInteruniversitario, the largest Italian computing centre (Italy)
CPU	Central Processing Unit
CRAC	Computer Room Air Conditioner
CRAH	Computer Room Air Handler
CSC	Finnish IT Centre for Science (Finland)
CSCS	The Swiss National Supercomputing Centre (represented in PRACE by ETHZ, Switzerland)
DC	Direct Current
DDR	Double Data Rate
DGAS	Distributed Global Address Space
DM	Distributed Memory
DNO	Distribution Network Operator (power/electricity supply)
DoW	Description of Work of PRACE project
DP	Double Precision, usually 64-bit floating point numbers
EC	European Commission
EFlop/s	Exa (= 10^{18}) Floating point operations (usually in 64-bit, i.e. DP) per second, also EF/s
EMEA	Europe, Middle-East & Africa (economic and business area)
EPCC	Edinburg Parallel Computing Centre (represented in PRACE by EPSRC, United Kingdom)
EPSRC	The Engineering and Physical Sciences Research Council (United Kingdom)

ETHZ	Eidgenössische Technische Hochschule Zürich, ETH Zurich (Switzerland)
EVB-IT	Ergänzende Vertragsbedingungen für die Beschaffung von IT-Leistungen (Additional conditionsofcontractfortheprocurementof IT services)
FDR	Fourteen Data Rate InfiniBand™ (14Gb/s data rate per lane)
Flop/s	Floating point operations (usually in 64-bit, i.e. DP - Double Precision)
FSA	Fusion System Architecture
FZJ	ForschungszentrumJülich (Germany)
GB	Giga (= 2^{30} ~10^9) Bytes (= 8 bits), also GByte
GCS	Gauss Centre for Supercomputing (Germany)
GENCI	Grand Equipement National de Calcul Intensif (France)
GFlop/s	Giga (= 10^9) Floating point operations (usually in 64-bit, i.e. DP) per second, also GF/s
GHz	Giga (= 10^9) Hertz, frequency =10^9 periods or clock cycles per second
GPFS	General Parallel File System (IBM)
GPGPU	General Purpose GPU
GPU	Graphic Processing Unit
HLRS	Höchstleistungsrechenzentrum Stuttgart, High Performance Computing Centre Stuttgart
HPC	High Performance Computing; Computing at the highest performance level at any given time; often-used synonym with Supercomputing
IBM	Formerly known as International Business Machines
I/O	Input/Output
KTH	KungligaTekniskaHögskolan (represented in PRACE by SNIC, Sweden)
LANL	Los AlamosNationalLaboratory, Los Alamos, New Mexico (USA)
LLNL	Laurence Livermore National Laboratory, Livermore, California (USA)
LRZ	Leibniz Supercomputing Centre (Garching, Germany)
LV	Low Voltage
MFlop/s	Mega (= 10^6) Floating point operations (usually in 64-bit, i.e. DP) per second, also MF/s
MHz	Mega (= 10^6) Hertz, frequency =10^6 periods or clock cycles per second
MPI	Message Passing Interface
MPICH	Freely available, portable implementation of MPI
MPP	Massively Parallel Processing (or Processor)
NICS	National Institute for Computational Sciences (NSF), Oak Ridge, Tennessee (USA)
NSF	National Science Foundation (USA)
OJEU	Official Journal of the European Union
OPENCL	Open Computing Language
ORNL	Oak Ridge National Laboratory, Oak Ridge, Tennessee (USA)
PCP	Pre-commercial procurement
PDU	Power Distribution Unit
PFlop/s	Peta (= 10^{15}) Floating-point operations (usually in 64-bit, i.e. DP) per second, also PF/s
PGAS	Partitioned Global Address Space
PRACE	Partnership for Advanced Computing in Europe; Project Acronym
PRACE-PP	PRACE Preparatory Phase (January 2008 – June 2010)
PRACE-1IP	PRACE 1st Implementation Phase (July 2010 – June 2012)
PSNC	Poznan Supercomputing and Networking Center (Poland)
QDR	Quad Data Rate

R&D	Research and Development
RBS	Risk Breakdown Structure
RFI	Request for Information
RfP	Request for Proposal
RI	(European) Research Infrastructure
RISC	Reduced Instruction Set Computing
ROI	Return On Investment
RSS	Resource Description Framework(RDF) Site Summary Really Simple Syndication, XML application conform to W3C RDF
SDK	Software Development Kit
SM	Shared Memory
SNIC	Swedish National Infrastructure for Computing (Sweden)
STFC	Science and Technology Facilities Council (represented in PRACE by EPSRC, United Kingdom)
TCO	Total Cost of Ownership.Includes the costs of personnel, power, cooling, maintenance, in addition to the purchase cost of a system.
TFlop/s	Tera (= 10^{12}) Floating-point operations (usually in 64-bit, i.e. DP) per second, also TF/s
TGCC	Très Grand Centre de Calcul du CEA
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
UPS	Uninterruptible Power Supply
URL	Uniform Resource Locator
WP7	PRACE Work Package 7 - Enabling Petascale Applications: Efficient Use of Tier-0 Systems (PRACE-2IP Project)
WP8	PRACE Work Package 8 - Support for the procurement and Commissioning of HPC service (PRACE-1IP Project)
WTO	World Trade Organisation
XPD	Heat exchanger units

Executive Summary

The PRACE-1IP Work Package 8 (WP8), “Support for the procurement and Commissioning of HPC service”, has three objectives:

- Assessment of petascale systems to be considered as PRACE production systems (task 1);
- Design and operation of reliable and power efficient high performance computing (HPC) computer centres (task 2);
- Sharing of best practices for the procurement and installation of HPC systems (task 3);

This Work Package builds on and expands the important work started in the PRACE Preparatory Phase project (PRACE-PP, January 2008 – June 2010), which sought to reach informed decisions within PRACE as a whole on the acquisition and hosting of HPC systems.

WP8 provides input for defining and updating the procurement plan and strategy and fosters the sharing of best practices for procuring and installing production systems. This deliverable(D8.4–Final Report on Petascale Systems and Centre Assessment) formalizes such input by summarizing the work achieved during WP8 (July 2010 – June 2012 period).

Task 1 - Assessment of petascale systems - has performed a continuous market watch and analysis of trends. The analysis also encompasses comparisons and graphs that allow an easier and quicker examination of trends for different aspects of top-level HPC systems. Work has also been carried out to automate and facilitate the gathering of information for the market watch. Web tools have been created for this, allowing anybody involved in PRACE to quickly search through the useful sources that were identified.

Hardware-software correlation has been explored by means of programming models, benchmarks and applications (using performance models), to assess the suitability of architectures for real-world applications, which is of key importance in assessing the real value of systems. This work is being carried out in collaboration with WP7 - Enabling Petascale Applications: Efficient Use of Tier-0 Systems”.

Task 2 - Design and operation of reliable and power efficient HPC computer centres - has produced a series ofwhite papers, each of which explores a specific topic related to HPC data centre design and operation. Dedicated surveys were circulated among PRACE sites to collect as much information as possible from the people involved in facility construction and maintenance, as well as managers. The results are analysed and assembled into a stand-alone document for each topic, for use as reference by PRACE partners.

A brief description of the “European Workshop on HPC Centre Infrastructures” series, which were hosted by CSCS, CEA then LRZ, is also included in the document.

Task 3 - Best practices for the procurement and installation of HPC systems - has proceeded with their plans to capture information and experience from PRACE partners, based on a synthetic questionnaire. The survey has been sent to many PRACE sites with recent procurements (a mix of Tier-1 and Tier-0) and their responses have been received, compiled, and analysed. Some introductory information has also been gathered about pre-commercial procurements in the field of HPC, a topic that will be of interest for the PRACE-3IP project. A white paper on HPC systems procurement has been produced from Task 3 material and findings.

1 Introduction

As stated in the Description of Work (DoW) of PRACE First Implementation Phase (1IP), the objectives of WP8 are the:

- Assessment of petascale systems to be considered as PRACE production systems;
- Design and operation of reliable and power efficient HPC computer centres;
- Sharing of best practices in procurements including related risk management.

This Work Package builds on and expands the important work started in the PRACE Preparatory Phase project, which sought to reach informed decisions within PRACE as a whole on the acquisition and hosting of HPC systems. WP8 provides input for defining and updating the procurement plan and strategy and fosters the sharing of best practices for procuring and installing production systems. It promotes the visibility of the RI amongst HPC vendors.

This document D8.4 - Final Report on Petascale Systems and Centre Assessment - is the last deliverable of a set of four deliverables to be produced by WP8. Its aim is to continue the work started in D8.1, D8.2 and D8.3 [26][27][28] – all three being PRACE internal reports - and summarise it all for public release.

D8.4 is organized in 5 main chapters, in addition to this introduction (Chapter 1) and to the conclusion (Chapter 7):

- An update on the HPC market and its evolution with different sources of information in Chapter 2;
- A discussion of hardware-software correlation vision, in relation with WP7 benchmarking and performance modelling approaches, in Chapter 3;
- A summary of whitepapers on HPC infrastructure topics in Chapter 4;
- A report on the European Workshops on HPC centre infrastructures in Chapter 5;
- A summary of surveys on recent procurement of large HPC systems in Chapter 6.

Some additional details are provided in annex:

- Surveys on HPC infrastructure topics used for related white papers of Chapter 4
- Programmes of European Workshops on HPC centre infrastructures mentioned in Chapter 5
- Questionnaire on recent procurements, related to Chapter 6

Several white papers on procurement and infrastructure related issues have been produced by WP8 [29][30][31][32][33][34]. They are being or will be made available and disseminated via the PRACE web site [35].

2 Assessment of petascale systems

The global competition for supercomputing capability is continuously expanding in performance and investment, and now more than ever before it is also expanding in international reach. The playing field is shaped by a number of technological and economic factors – levels of investments in IT and HPC, technological breakthroughs that can push performance, mass market trends that can pull technologies. In this landscape, more countries are presenting their candidacy for the coveted top spot in worldwide supercomputing performance. Keeping a watch on this competition, marked by the bi-annual Top500 list of the most powerful supercomputers in the world, has become of the utmost importance for establishing strategies for advancement in such a hard-fought industry. It is therefore obvious that PRACE, in its mission of enhancing European competitiveness in scientific discovery and engineering research and development, is in need of this constant market watch.

Observing leading HPC systems around the globe provides a very good insight into the state of the market and technologies involved, and the deeper the examination goes the more useful conclusions can be extracted. By sorting and analysing the raw data, and comparing it periodically to add the time component, information is provided on the evolution of HPC in general, as well as specific details on technologies and other influencing factors. This chapter concentrates on presenting the information that has been collected concerning worldwide petascale systems and initiatives, and analysing it for the benefit of PRACE and its members.

The chapter is divided into two sections:

Market Watch summarises the information gathered for the analysis, providing a glimpse of the raw data and the means by which it was compiled.

Market Analysis describes the actual study performed, based on this information, and draws conclusions from the results of this study.

2.1 Market watch

This section presents a snapshot view of the current state of leading petascale HPC systems around the world. It includes details on production systems (installed and running) as well as procured and planned systems that will supposedly enter production in the next 2-3 years. The complete raw data collected for this Market Watch is available to PRACE partners in an internal wiki.

NB: as of writing this section, June 2012 Top 500 list had not been published yet.

2.1.1 *Snapshot*

Current systems

As of May 2012, there are at least twenty-one HPC systems in production that can be considered as "petascale" because their peak performance is around 1 Petaflop/s or more. Eighteen of these systems have proven their petascale capabilities by running the LINPACK benchmark for recognition in the November 2011 Top500 list, while the others are expected to join them next month with the publication of the June 2012 edition of Top500. In the subsequent analysis, only the eighteen ranked systems will be used for statistical comparison (since the data can be considered official), but the others will be mentioned in the remarks.

Table 1below gives a brief description of each of the current petascale production systems:

System	Site (Country)	Model (Processor)	LINPACK / peak (PFlop/s)
K Computer	RIKEN (Japan)	Fujitsu Cluster (Fujitsu SPARC64)	10.51 / **11.28**
Tianhe-1A	NSCT (China)	NUDT YH MPP (Intel Xeon / NVIDIA Tesla)	2.57 / **4.70**
Nebulae	NSCS (China)	Dawning TC3600 (Intel Xeon / NVIDIA Tesla)	1.27 / **2.98**
Jaguar	ORNL (USA)	Cray XT5/XT4 (AMD Opteron)	1.76 / **2.33**
Tsubame 2.0	GSIC-TIT (Japan)	NEC/HP ProLiant SL390s G7 (Intel Xeon / NVIDIA Tesla)	1.19 / **2.29**
Curie	TGCC (France)	Bull S6010/S6030 (Intel Xeon / NVIDIA Tesla)	N/A / **2.00**
Helios	IFERC (Japan)	Bull B510 (Intel Xeon)	N/A / **1.50**
Roadrunner	LANL (USA)	IBM TriBlade (AMD Opteron / IBM PowerXCell)	1.04 / **1.38**
Lomonosov	RCC (Russia)	T-Platforms T-Blade2/T-Blade1.1 (Intel Xeon / NVIDIA Tesla / IBM PowerXCell)	0.67 / **1.37**
Cielo	LANL (USA)	Cray XE6 (AMD Opteron)	1.11 / **1.37**
Tianhe-1A Hunan Solution	NSCH (China)	NUDT YH MPP (Intel Xeon / NVIDIA Tesla)	0.77 / **1.34**
Pleiades	NAS (USA)	SGI Altix ICE 8400EX/8200EX (Intel Xeon)	1.09 / **1.32**
Hopper	NERSC (USA)	Cray XE6 (AMD Opteron)	1.05 / **1.29**
Tera-100	CEA (France)	Bull S6010/S6030 (Intel Xeon)	1.05 / **1.26**
Kraken XT5	NICS-UT (USA)	Cray XT5 (AMD Opteron)	0.92 / **1.17**
Oakleaf-FX	SCD/ITC (Japan)	Fujitsu PRIMEHPC FX10 (Fujitsu SPARC64)	N/A / **1.13**
Sunway Blue Light	NSCJ (China)	NRCPCET Sunway BlueLight MPP (ShenWei SW1600)	0.80 / **1.07**
Hermit	HLRS (Germany)	Cray XE6 (AMD Opteron)	0.83 / **1.04**
Mole-8.5	IPE-CAS (China)	Tyan FT72-B7015 (Intel Xeon / NVIDIA Tesla)	0.50 / **1.01**
JUGENE	FZJ (Germany)	IBM Blue Gene/P (IBM PowerPC)	0.83 / **1.00**
Zin	LLNL (USA)	ApproXtreme-X GreenBlade GB512X (Intel Xeon)	0.77 / **0.96**

Table 1: Current petascale production systems

Next systems: near existing and future petascale supercomputers

Information on future systems is more limited and less trustworthy than that of production systems, and depends greatly on the stage of procurement of the computer. There are currently fourteen publicly known systems that will probably enter production in the next two years (most are scheduled for this year 2012) with a peak performance around or above 1 Petaflop/s. In addition to those procurements, there are two systems that are already in production and will undergo important updates in the coming years: Pleiades and Jaguar will become Carlsbad and Titan, respectively.

As of writing this document, European FERMI and SUPERMUC systems are installed and will enter production in the next weeks.

Only two of the announced supercomputers have an expected peak performance superior to the current leader, K Computer. Four of the systems will have a similar performance (around 10-11 Petaflop/s), while the other eight are in the 1-3 Petaflop/s range and probably won't reach the top 5 positions. Obviously these fourteen systems do not represent all the petascale systems that will enter production in the following 2 years, since many projects are more confidential and others might be in final stages of procurement but not yet announced. It is highly expected, for example, that China release more than one petascale system in 2012-2013.

Table 2summarises the details of these procurements:

System	Expected delivery	Site (Country)	Model (Processor)	Peak (PFlop/s)
Sequoia	2012	LLNL (USA)	IBM Blue Gene/Q (IBM PowerPC)	**20.13**
Titan (Jaguar update)	2012	ORNL (USA)	Cray XE6/XK6 (AMD Opteron / NVIDIA Tesla)	**20**
Blue Waters	2012	NCSA (USA)	Cray XE6/XK6 (AMD Opteron / NVIDIA Tesla)	**11.5**
Carlsbad 3.0 (Pleiades update)	2012	NAS (USA)	SGI Altix ICE (Intel Xeon)	**10**
Mira	2013	ANL (USA)	IBM Blue Gene/Q (IBM PowerPC)	**10**
Stampede	2013	TACC (USA)	Dell Zeus (Intel Xeon / Intel MIC)	**10**
SuperMUC	2012	LRZ (Germany)	IBM System x iDataPlex (Intel Xeon)	**3**
N/A	2012	CSTJF (France)	SGI Altix ICE X	**2.3**
Fermi	2012	CINECA (Italy)	IBM Blue Gene/Q (IBM PowerPC)	**2**
Yellowstone	2012	NCAR (USA)	IBM iDataPlex (Intel Xeon)	**1.6**
N/A	2012	EPCC (UK)	IBM Blue Gene/Q	**1.26**
Gaea	2012	NOAA (USA)	Cray XT6/XE6 (AMD Opteron)	**1.1**
MareNostrum 3	2012	BSC (Spain)	Unknown	**1**
Dawning 6000	2012	ICT-CAS (China)	Dawning (Loongson 3B)	**1**

Table 2: Next and future petascale systems

2.1.2 *Sources*

The aim of the following sections is to identify a number of useful and reliable web references that can serve as sources of information for the continuous market watch of PRACE-1IP WP8, as well as other stakeholders that might want to easily find insights about the current and future trends of the HPC market.

At the end of the section we document the developments that help the identification of information relevant to the market watch, namely the internal Market watch web page that contains links to the Netvibes feeds aggregator and the Market Watch Google Custom Search Engine (CSE).

We can identify four types of such sources on the web:

1. HPC related electronic publications / web sites: These publications facilitate the identification of news and opinions of various HPC experts, on a variety of subjects related to the HPC market, ranging from new technologies available from vendors, to new or expected purchases from computing centres around the world, to technology trends driven by vendors or by user demand. These web sites aggregate news from various sources and present both the vendors' as well as the users' views of the HPC market.
2. The web site of the computing centre hosting a supercomputer: These web sites contain the details about the supercomputers both on the technical equipment level as well as the intended usage.
3. Vendor specific web sites: These web sites, usually the main vendor web sites, contain a variety of information on the new technologies developed and deployed by them. These as presented in the form of product documentation, white papers, press releases etc. Further to that, on the vendor web sites one can find information on the collaborations and sales that a vendor has achieved through the press releases that the vendors issues. The vendor specific web sites essentially offer the vendor's point of view on the HPC market.
4. Funding agencies web sites: These web sites are maintained by various funding agencies around the world. This is where one can find information on new or planned procurements via press releases or RFIs/RFPs (Requests for Information/Proposal) that might be public.

Further to that we can categorize the web references based also on the categorization of the HPC systems that is followed throughout this deliverable, i.e.:

- Web sites containing information on existing systems;
- Web sites containing information on procured systems;
- Web site containing information on planned systems.

The following sections provide a simple list of relevant web references based on the four categorizations that were described above.

HPC Related Electronic Publications / Web Sites

Top500	http://www.hpcwire.com
Green500	http://www.green500.org
HPC Wire	http://www.top500.org
HPC Inside	http://insidehpc.com/
HPC Projects	http://www.hpcprojects.com/
Scientific Computing .COM	http://www.scientific-computing.com/
Microprocessor report	http://www.mdronline.com/
Supercomputing online	http://www.supercomputingonline.com/

Table 3: Web sites for searching for information on current and future HPC systems

Table 3 above lists a number of useful web sites for searching for information on current and future HPC systems, some of them commented below.

For current systems:

- http://www.top500.org – Top 500 supercomputer sites

The Top500 supercomputer site publishes the Top500 list of general purpose systems that are in common use for HPC applications. The present Top500 list lists computers ranked by their performance on the LINPACK Benchmark. The list is updated half-yearly, keeping track of the evolution of computers.

- http://www.green500.org – The Green 500

The purpose of the Green500 is to provide a ranking of the most energy-efficient supercomputers in the world. In order to raise awareness to other performance metrics of interest (e.g., performance per watt and energy efficiency for improved reliability), the Green500 offers lists to encourage supercomputing stakeholders to ensure that supercomputers are only simulating climate change and not creating climate change. The list is updated half-yearly and it uses as ranking metric the "Flop/s-per-Watt.

For current and future systems:

- http://www.hpcwire.com – HPC Wire

This is an on line publication devoted to HPC news. It is one of the most popular online publications for people involved in High Performance Computing. Newsis categorized into several topics, such as: Applications, Developer Tools, Interconnects, Middleware, Networks, Processors, Storage, Systems and Visualization. Special sections exist for the different industries that are related to HPC, such as: Academia & Research, Financial Services, Government, Life Sciences, Manufacturing, Oil & Gas and Retail.

- In this category we can also include the International Exascale Software Project - http://www.exascale.org/.

The goal of the IESP is to come up with an international plan for developing the next generation of open source software for high performance scientific computing. The project organises meetings and produces documents where experts and leading HPC players around the world present, amongst other things, details on existing systems and/or their plans for future exascale systems (i.e. targeting the Exaflop/s). The documents and presentations from the meetings are publicly available and constitute a very good source of information for the market watch.

- In this category we can also add the EESI (European Exascale Software Initiative) - http://www.eesi-project.eu

The objective of this Support Action, co-funded by the European Commission is to build a European vision and roadmap to address the programming and application challenges of the new generation of massively parallel systems composed of millions of heterogeneous cores - frompetascale in 2010 to foreseen exascale performances in 2020.

The documents and presentations from the meetings are publicly available and constitute a very good source of information for the market watch.

Companies such as IDC or GARTNER also are sources of valuable market information and have a special focus on HPC activities. Their offer is mostly commercial but there is some public dissemination of select and synthetic information (e.g., IDC and HPC User Forum http://www.hpcuserforum.com/, or regular market updates with some predictions and forecast).

General reference:

IDC	http://www.idc.com/
GARTNER	http://www.gartner.com/

Computing centre web sites

The list of computing centre web sites can be obtained from the Top500 list. For completeness of this deliverable we list in Table 4 the Top 10 supercomputing sites as of November 2011 and their web sites.

Rank	Site Name	Computer	Web Address
1	RIKEN Advanced Institute for Computational Science (AICS)	K computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect	http://www.aics.riken.jp/en/kcomputer/
2	National Supercomputing Center in Tianjin	Tianhe-1A - NUDT TH MPP, X5670 2.93Ghz 6C, NVIDIA GPU, FT-1000 8C	http://www.nscc-tj.gov.cn/en/ http://www.nscc-tj.gov.cn/en/show.asp?id=191
3	DOE/SC/Oak Ridge National Laboratory	Jaguar - Cray XT5-HE Opteron 6-core 2.6 GHz	http://computing.ornl.gov/ http://www.nccs.gov/jaguar/
4	National Supercomputing Centre in Shenzhen (NSCS)	Nebulae - Dawning TC3600 Blade, Intel X5650, NVidia Tesla C2050 GPU	N/A
5	GSIC Center, Tokyo Institute of Technology	TSUBAME 2.0 - HP ProLiant SL390s G7 Xeon 6C X5670, Nvidia GPU, Linux/Windows	http://www.gsic.titech.ac.jp/ http://tsubame.gsic.titech.ac.jp/en

6	DOE/NNSA/LANL/SNL	Cielo - Cray XE6 8-core 2.4 GHz	http://www.lanl.gov/orgs/hpc/cielo/index.shtml
7	NASA/Ames Research Center/NAS	Pleiades - SGI Altix ICE 8200EX/8400EX, Xeon HT QC 3.0/Xeon 5570/5670 2.93 Ghz, Infiniband	http://www.nas.nasa.gov/hecc/resources/pleiades.html
8	DOE/SC/LBNL/NERSC	Hopper- Cray XE6 12-core 2.1 GHz	http://www.nersc.gov/ http://www.nersc.gov/nusers/systems/hopper2/
9	CEA Commissariat à l'Energie Atomique et aux Energies Alternatives	Tera-100 - Bull bullx super-node S6010/S6030	http://www.cea.fr/ http://www-hpc.cea.fr/fr/complexe/docs/T100.htm
10	DOE/NNSA/LANL	Roadrunner - BladeCenter QS22/LS21 Cluster, PowerXCell 8i 3.2 Ghz / Opteron DC 1.8 GHz, Voltaire Infiniband	http://www.lanl.gov/ http://www.lanl.gov/roadrunner/

Table 4: Web sites for the 10 most powerful systems in the Top500 (November 2011)

Vendors Web Sites

There are a large number of companies that design and produce HPC related hardware and software. For example, during the latest Supercomputing Conference (SC11) in New Orleans, 207 commercial exhibitors showed their products.

The following list of vendors in Table 5 is based on the vendors that supplied the most powerful 50 systems of the November 2011 Top500 list. Note that National University of Defence Technology (NUDT) and the Institute of Processing Engineering, of the Chinese Academy of Sciences (IPE), are not included since they are institutes and cannot be considered as global vendors.

Vendor Name	HPC Related Web Pages	URL of Press Releases or RSS feed
Appro International	http://www.appro.com/	http://www.appro.com/press/press.asp
Bull SA	http://www.bull.com/extreme-computing/index.html	http://www.bull.com/about-bull/news.html
ClusterVision	http://www.clustervision.com/products	http://www.clustervision.com/news
Cray Inc.	http://www.cray.com/Products/Products.aspx	http://www.cray.com/About/Newsroom.aspx
Dawning	http://www.dawning.com.cn/	

Vendor Name	HPC Related Web Pages	URL of Press Releases or RSS feed
Dell	http://content.dell.com/us/en/enterprise/hpcc.aspx?cs=555	http://content-cdn.dell.com/api/feeds.svc/rss/newsroom-press-releases?c=us&l=en&s=corp&cs
Fujitsu	http://www.fujitsu.com/global/services/solutions/tc/hpc/products/ http://www.fujitsu.com/global/services/solutions/tc/supercomputer/	http://www.fujitsu.com/us/news/pr/recent/index_computing.html
Hewlett-Packard	http://h20311.www2.hp.com/hpc/us/en/hpc-index.html	http://www.hp.com/hpinfo/newsroom/index.html?mtxs=corp&mtxb=3&mtxl=4
IBM	http://www-03.ibm.com/systems/technicalcomputing	http://www-03.ibm.com/systems/technicalcomputing/news
NEC	http://www.nec.com/en/de/en/prod/servers/hpc/index.html	http://www.necam.com/About/Press_Center.cfm
NVIDIA	http://www.nvidia.com/object/tesla_computing_solutions.html	http://nvidianews.nvidia.com/ http://www.nvidia.com/object/rss_home.html
SGI	http://www.sgi.com/	http://www.sgi.com/company_info/newsroom/press_releases/2010/index.html
Oracle	http://www.oracle.com/us/products/servers-storage/index.html	http://www.oracle.com/us/corporate/press/index.html http://www.oracle.com/us/corporate/press/rss/rss-pr.xml
Raytheon	http://www.raytheon.com/capabilities/products/hpc/	http://raytheon.mediaroom.com/index.php?year=2011&s=43
SuperMicro	http://www.supermicro.com/index.cfm	http://www.supermicro.com/newsroom/pressreleases/
T-Platforms	http://www.t-platforms.ru/	http://www.t-platforms.ru/about-company/press-releases
TYAN	http://www.tyan.com/products.aspx	http://www.tyan.com/newsroom_pressroom.aspx

Table 5: Websites from HPC vendors

Funding Agencies Web Sites

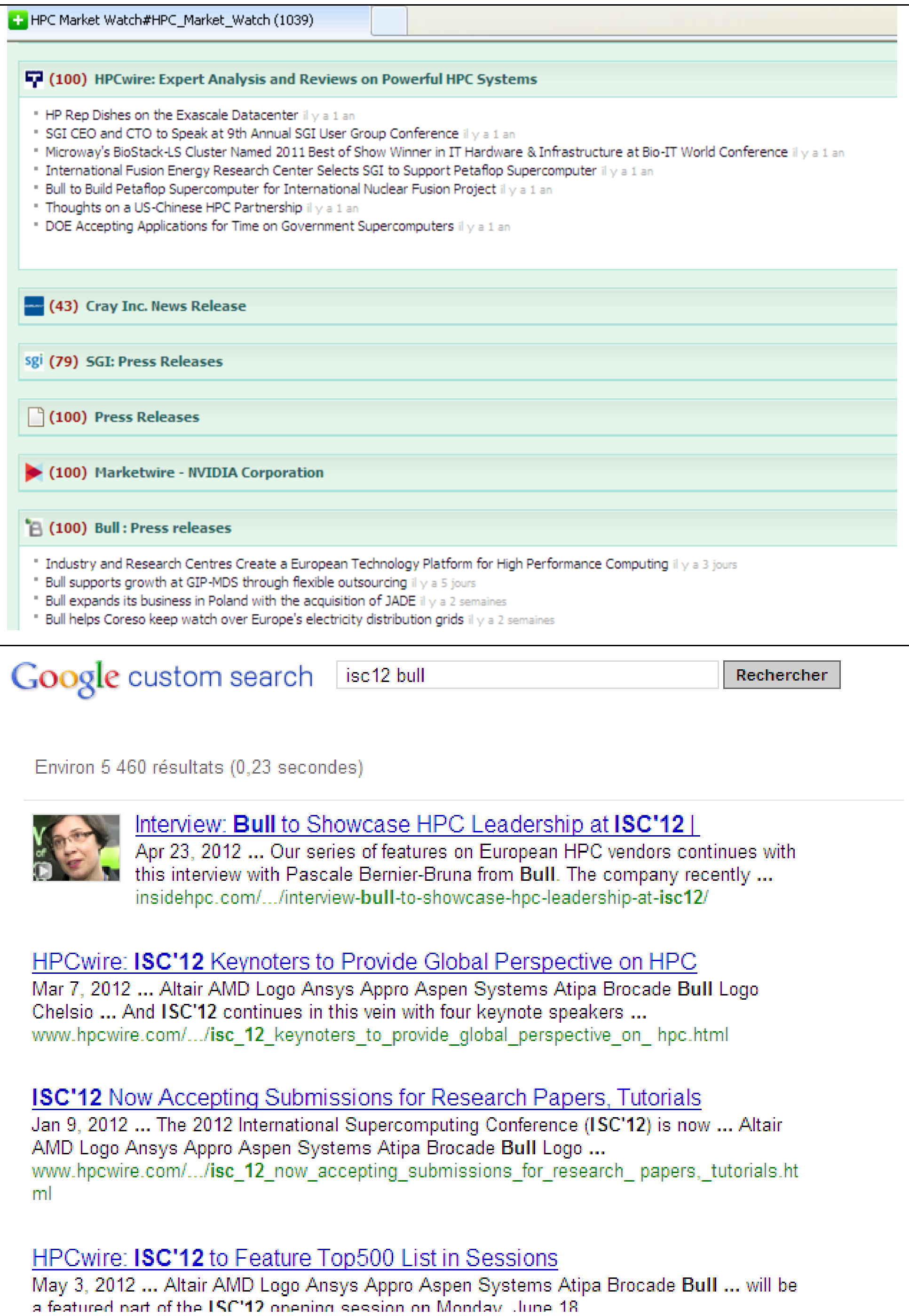
The following Table 6 presents the URLs of major funding bodies outside Europe. For funding available within Europe, PRACE is informed by the participating institutes and partners.

Country	Agency	URL
USA	Department of Energy (DOE), Advanced Scientific Computing Research (ASCR)	http://www.science.doe.gov/ascr/News/News.html http://science.energy.gov/news/
USA	Department of Energy (DOE), National Nuclear Security Administration	http://nnsa.energy.gov/
USA	Department of Defense (DOD)	http://www.hpcmo.hpc.mil/cms2/index.php
USA	Department of Defense (DOD), Defense Advanced Research Projects Agency (DARPA)	http://www.darpa.mil/
USA	NASA, Ames Exploration Technology Directorate	http://infotech.arc.nasa.gov/
USA	National Science Foundation, CyberInfrastructure (OCI)	http://www.nsf.gov/dir/index.jsp?org=OCI
USA	National Nuclear Security Administration (NNSA)	http://nnsa.energy.gov/
Japan	Council for Science and Technology Policy (CSTP)	http://www8.cao.go.jp/cstp/english/index.html
Japan	Japan Science and Technology Agency (JST)	http://www.jst.go.jp/EN/
Japan	Ministry of education, culture, sport, science and technology (MEXT)	http://www.mext.go.jp/english/
China	Ministry of Science and Technology of the People's republic of China	http://www.most.gov.cn/eng/
China	National Natural Science Foundation of China (NSFC)	http://www.nsfc.gov.cn/e_nsfc/desktop/zn/0101.htm
South Korea	Korean Ministry of Education, Science and Technology (MEST)	http://english.mest.go.kr/enMain.do
India	Planning Commission of the government of India. Science and Technology section.	http://planningcommission.nic.in/sectors/index.php?sectors=sci
Russia	Rosatom State Nuclear Energy Corporation	http://www.rosatom.ru/en/

Table 6: Funding agencies web sites

Market Watch Tools

Two tools have been deployed within PRACE in order to take advantage of the above collection of links for the Market watch. These tools facilitate the aggregation of the links (where possible) to a single web page and the creation of a Google custom search engine that allows for search queries within a pre-defined set of URLs (Figure 1). Both tools as well as an up to date list of the web sources that appeared in the previous sections are available to WP8 via a PRACE internal wiki page.The tools were mainly used internally by the WP8 team in order to monitor the HPC market developments and trends.



HPC Market Watch#HPC_Market_Watch (1039)

(100) HPCwire: Expert Analysis and Reviews on Powerful HPC Systems

- HP Rep Dishes on the Exascale Datacenter il y a 1 an
- SGI CEO and CTO to Speak at 9th Annual SGI User Group Conference il y a 1 an
- Microway's BioStack-LS Cluster Named 2011 Best of Show Winner in IT Hardware & Infrastructure at Bio-IT World Conference il y a 1 an
- International Fusion Energy Research Center Selects SGI to Support Petaflop Supercomputer il y a 1 an
- Bull to Build Petaflop Supercomputer for International Nuclear Fusion Project il y a 1 an
- Thoughts on a US-Chinese HPC Partnership il y a 1 an
- DOE Accepting Applications for Time on Government Supercomputers il y a 1 an

(43) Cray Inc. News Release

(79) SGI: Press Releases

(100) Press Releases

(100) Marketwire - NVIDIA Corporation

(100) Bull : Press releases

- Industry and Research Centres Create a European Technology Platform for High Performance Computing il y a 3 jours
- Bull supports growth at GIP-MDS through flexible outsourcing il y a 5 jours
- Bull expands its business in Poland with the acquisition of JADE il y a 2 semaines
- Bull helps Coreso keep watch over Europe's electricity distribution grids il y a 2 semaines

Google custom search isc12 bull Rechercher

Environ 5 460 résultats (0,23 secondes)

Interview: **Bull** to Showcase HPC Leadership at **ISC'12** |
Apr 23, 2012 ... Our series of features on European HPC vendors continues with this interview with Pascale Bernier-Bruna from **Bull**. The company recently ...
insidehpc.com/.../interview-**bull**-to-showcase-hpc-leadership-at-**isc12**/

HPCwire: **ISC'12** Keynoters to Provide Global Perspective on HPC
Mar 7, 2012 ... Altair AMD Logo Ansys Appro Aspen Systems Atipa Brocade **Bull** Logo Chelsio ... And **ISC'12** continues in this vein with four keynote speakers ...
www.hpcwire.com/.../**isc_12**_keynoters_to_provide_global_perspective_on_ hpc.html

ISC'12 Now Accepting Submissions for Research Papers, Tutorials
Jan 9, 2012 ... The 2012 International Supercomputing Conference (**ISC'12**) is now ... Altair AMD Logo Ansys Appro Aspen Systems Atipa Brocade **Bull** Logo ...
www.hpcwire.com/.../**isc_12**_now_accepting_submissions_for_research_ papers,_tutorials.html

HPCwire: **ISC'12** to Feature Top500 List in Sessions
May 3, 2012 ... Altair AMD Logo Ansys Appro Aspen Systems Atipa Brocade **Bull** ... will be a featured part of the **ISC'12** opening session on Monday, June 18

Figure 1: Customised web search tools / netvibes aggregator, google custom search

Feeds aggregators and Netvibes

A feed aggregator is a software package or a Web application which aggregates syndicated web content such as RSS feeds, blogs, social networks content etc. in a single location for easy viewing. For the purposes of the Market Watch activity of WP8 the Netvibes[1] aggregator has been used. Netvibes is a personalized dashboard publishing platform for the Web including digital life management, widget distribution services and brand observation rooms. It is organized into tabs, with each tab containing user-defined modules. Built-in Netvibes modules include an RSS/Atom feed reader, social network feed readers and a big variety of other online service's readers. For the purposes of the Market Watch we used the RSS and twitter readers. Twitter accounts where used for some of the vendor's sites whose news web pages were not available via RSS feeds. The page can be accessed at [3].

Google Custom Search Engine

To facilitate a more efficient search among the results of Google searches we created an HPC Market Watch Google Custom Search Engine (CSE). CSE allows the creation of customised search engine using the Google search, by limiting the search space to only a predefined set of web sites[2]. That way CSE provides only relevant search results speeding the process of searching information. Within WP8 we have created a Market Watch CSE that contains 48 sites that are relevant to the activity. The CSE can be accessed at [4]. It can be used like a plain Google search.

2.2 Market analysis

This section provides an analysis of the current and near-future HPC ecosystem, based on the information contained in the market watch. Through the observation and comparison of the data, including the use of statistics and graphs, relevant conclusions are made regarding the situation and prospects of petascale systems around the globe. It must be noticed however, that the studied sets are quite small and that some related statistics must be considered as mere indications and trends.

The analysis is organized in two parts:

A static analysis draws conclusions exclusively from the examination of the current version of the market watch. It provides a detailed evaluation of the present situation, identifying the most popular features in petascale supercomputers. Statistically analysing each aspect of the current petascale systems reveals patterns that can help draw conclusions on the state of leading HPC systems.

A dynamic analysis compares current data and conclusions with those made in previous analyses [26][27][28], trying to detect trends and tendencies. It follows the evolution of system performance as well as architecture features.

2.2.1 *Static analysis*

Although the performance of 1 Petaflop/s was reached for the first time in 2008, all of the current petascale systems were either built or updated in 2009 (17%), 2010 (28%), and mainly 2011 (55%). Japan has taken the lead position based on performance, but the USA still clearly dominates in terms of market share in number of systems, with seven of the eighteen systems (39%,Figure 2). China's five machines give them a 28% share, while Germany and Japan tie in third place with two systems each (11%). France and Russia each have one system (5.5%).

As was mentioned in the Market Watch, both France and Japan have an additional production petascale system waiting to be officially benchmarked next month, which should increase their market share slightly, unless the USA and China manage to have some of their planned systems included in the next Top500 as well.

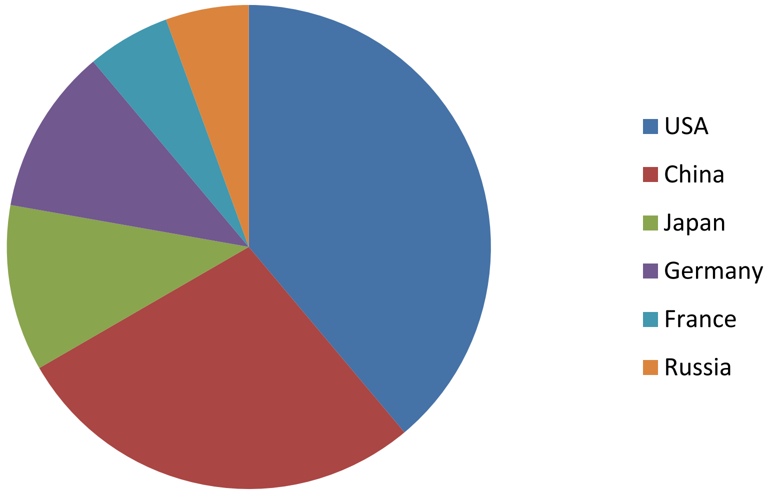


Figure 2: Petascale systems locations

The only site which has more than one petascale computer is Los Alamos National Laboratory (LANL), in New Mexico (USA), with both Roadrunner and Cielo. The city of Oak Ridge (Tennessee, USA) also has two petascale systems, but despite sharing the same location the two systems are owned by different organizations: Jaguar belongs to Oak Ridge National Laboratory (ORNL), funded by the Department of Energy (DOE) whilst the Kraken XT5 is part of the National Institute for Computational Sciences (NICS) which is funded by the National Science Foundation (NSF).

It is complicated to extract firm conclusions regarding the cost of petascale systems, since some funding agencies and sites regard such information as confidential, and in many cases the funding is part of a larger project, where the procurement of the computer is only a part of the total cost. Of the eighteen systems analysed, only seven provided a public figure in relation to cost. Of these, the most expensive is Jaguar at 160 million euro, but this includes the entire lifetime of the system since 2005 (upgraded from XT3 to XT4 and then to XT5). The cheapest petascale machine is Tsubame 2.0, which cost approximately 25 million euro. The average price of a petascale system, according to the available data, is around 70 million Euros. Unfortunately, all these figures must be taken only as very rough values, for the reasons previously indicated.

Static analysis - Performance

Peak performance of the analysed systems ranges from the 0.96 Petaflop/s of Zin, installed at LLNL, to the 11.28 Petaflop/s of RIKEN's K Computer. The total peak performance of the eighteen computers is almost 40 Petaflop/s, producing an average peak performance of 2.17 Petaflop/s per machine. The new systems that are expected to enter the June Top500 list are all in this same performance range, falling mostly around the average 2 Petaflop/s.

Actual performance as measured by the LINPACK benchmark is considerably lower than peak performance in some cases. The best result is again obtained by K Computer (Figure 3), and with a very high relative value: 10.51 Petaflop/s. The lowest performance on the

LINPACK benchmark does not correspond with the lowest peak performer but with the Chinese Mole-8.5 system, which uses GPU accelerators, with 0.5 Petaflop/s. On average the systems achieved 1.61 Petaflop/s executing the benchmark, approximately 74% of the mean peak performance.

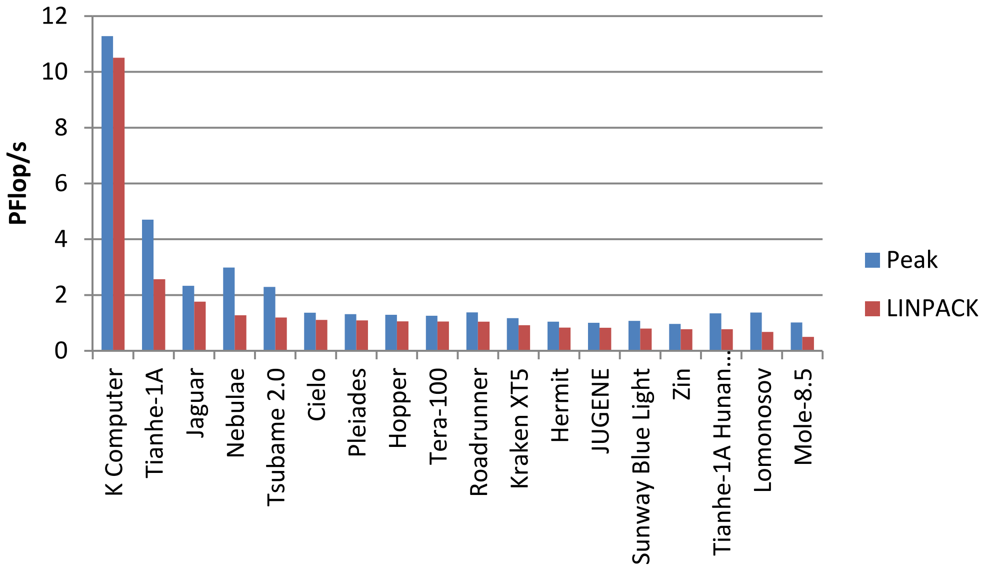


Figure 3: Petascale systems performance

Static analysis - Components

The most represented vendor within these petascale systems is Cray Inc. (Figure 4), who produced five of them (28% market share). IBM is the second most popular commercial petascale vendor, with two systems providing them with a 11% share. The National University of Defense Technology (NUDT) of China, although not a commercial vendor, also has two petascale systems and therefore controls 11% of the market. The situation is similar with the National Research Centre of Parallel Computer Engineering & Technology (NRCPCET), also a Chinese non-commercial vendor, responsible for the construction of the Sunway Blue Light. The rest of the vendors all have one machine each: Dawning, Bull, SGI, Fujitsu, Appro, T-Platforms, Tyan, and an NEC/HP consortium.

Fujitsu is both set to have a second system included in this list by next month (Oakleaf-FX) and Bull two more ones (Helios and Curie, respectively).

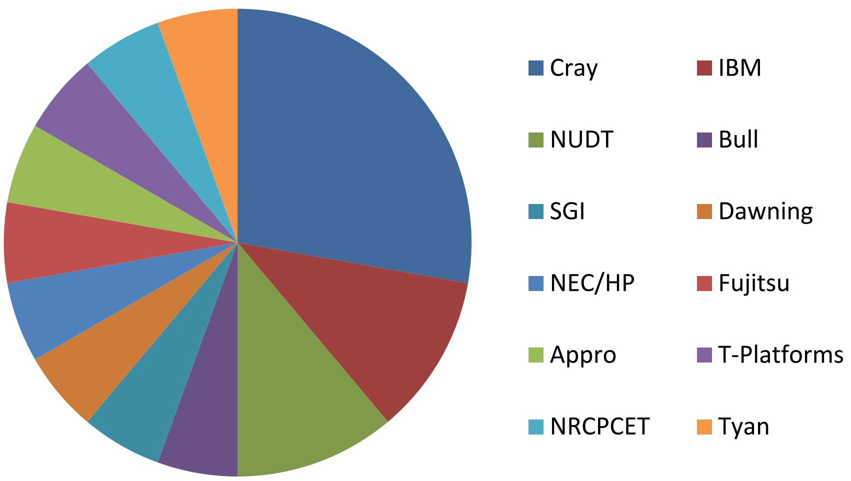


Figure 4: Petascale systems vendors

The most popular model for reaching petascale performance is the Cray XE6, selected for three of the production systems. Cray's older XT5 model is also used in two petascale computers. Other commercially available models that have achieved petascale are: Dawning TC3600, HP ProLiant SL390s G7, Bull S6010/S6030, SGIAltix ICE 8400EX/8200EX, T-Platforms T-Blade2, ApproXtreme-X GreenBlade, Sunway BlueLight MPP, IBM Blue Gene/P, and Tyan FT72-B7015. Three systems made use of non-standard experimental models: NUDT YH MPP (both Tianhe-1A variants) and IBM TriBlade (Roadrunner).

Intel and AMD dominate the processor market in general, and petascale computers are not an exception (Figure 5). Versions of their Xeon and Opteron series of processors are present in fifteen of the eighteen systems, with Intel leading AMD by three systems (50% vs. 33%). Another traditional player, IBM, uses its PowerPC processor in their Blue Gene/P supercomputer (JUGENE). The new players in the HPC processor market are Japanese company Fujitsu, with their SPARC64 VIIIfx processor, and Chinese company ShenWei, with their SW1600. It is already certain that Fujitsu will be adding at least one more petascale system based on their SPARC64 architecture (in this case the newer IXfx model) with Oakleaf-FX, while Intel will elevate their leader's share with Curie.

Clock frequencies range from the 850 MHz of the PowerPC 450 used in JUGENE to the 3 GHz of the Intel Xeon E5472 of Pleiades, averaging 2.37 GHz. Considering that there is a difference of 2.15 GHz between the most and least powerful processors, the average is much closer to the maximum value (3 GHz) than the minimum, showing that most systems use “high-power” processors.

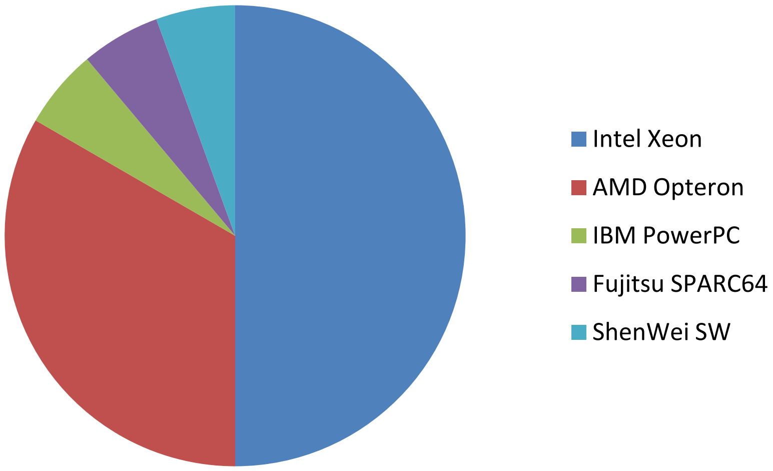


Figure 5: Processors used in petascale systems

Although the use of accelerator units seems to be on the rise, the majority (61%) of the systems in this analysis do not use any such device (Figure 6). Out of the seven computers integrating accelerators, five of them (accounting for 71%, or 28% of the total) use NVIDIA Tesla C2050, one uses the IBM PowerXCell 8i, and one (Lomonosov) uses both NVIDIA Tesla C2070 and IBM PowerXCell 8i.

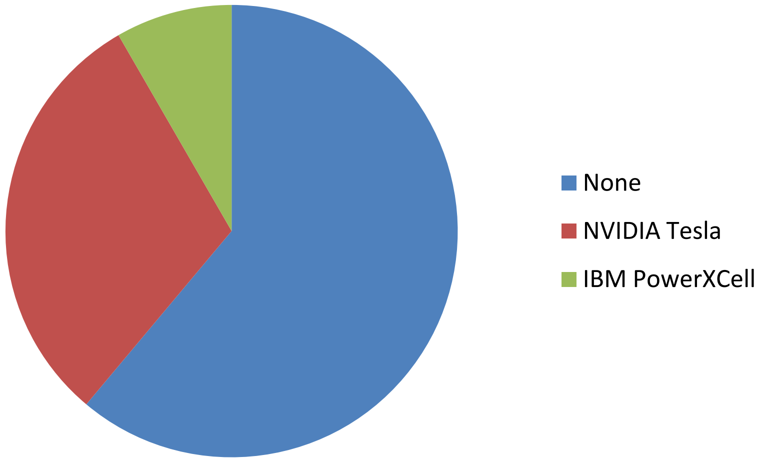


Figure 6: Accelerators used in petascale systems

System memory ranges between 1,377 TB (K Computer) and 15 TB (Mole-8.5), with an average of 238 TB per machine.

The number of CPU cores is highly disparate among the machines, ranging from as low as 29,440 for Mole-8.5 to K Computer's 705,024 (the average number of cores being 145,355). One reason for this large difference is that accelerator cores are not taken into account, meaning that machines that depend strongly on accelerators have a very small count. Adding accelerator cores and CPU cores can be deceptive, since accelerator cores are not clearly defined and are definitely not equivalent to CPU cores. If we analyse accelerated and non-accelerated systems separately, we can see that the CPU core count of the former is between 33,072 and 86,016 (average of 60,259 cores), while in the case of the non-accelerator computers the range is from 46,208 to 705,024 (196,659 cores on average). This is a clear indication that CPU core count cannot be compared fairly between these different types of petascale systems without distinguishing them beforehand.

Nodes also vary considerably between the systems, from 288 (Mole-8.5) to 88,128 (K Computer), which is equivalent to an average of 14,933 nodes. Interestingly, fifteen of the eighteen systems (83.33%) had a lower number of nodes than the average. The reason for this is that the three that have more nodes than average have many more: 88,128 (K Computer), 73,728 (JUGENE), and 26,520 (Jaguar).

A total of seven interconnect technologies are used in the eighteen petascale systems in this analysis (Figure 7) which are described in Table 7below:

Name	Owner	Description
InfiniBand DDR	The InfiniBand Trade Association (Industry Standard)	Double data rate (DDR) InfiniBand has a signalling rate of 5 Gbit/s, which effectively provides 4 Gbit/s per link (implementers can choose between 1, 4, and 12 links). Switch chips have a latency of 140 ns, working with 4 KB messages.
Gemini	Intel (originally Cray)	Two pairs of Opteron chips linking into the Gemini interconnect chip using HT3 links. The Gemini chip has 48 ports that have an aggregate bandwidth of 168 GB/s, and takes just a hair above one microsecond to jump between computer nodes hooked to different Gemini chips, and less than one microsecond to jump from any of the four processors talking to the same Gemini.
SeaStar 2+	Intel (originally Cray)	The SeaStar interconnect takes one HT link coming off a pair of Opteron processors and hooks it into a six-port router, with each port able to deliver 9.6 GB/sec of bandwidth. The HT2 links provide 25.6 GB/sec of bandwidth to memory on the two sockets of Opterons, 6.4 GB/sec of bandwidth from the processors out to the SeaStar2+ ASIC, and then six router ports running at 9.6 GB/sec each, implementing a 3D torus.
Arch	NUDT	The switch at the heart of Arch has a bi-directional bandwidth of 160 Gb/sec, a latency for a node hop of 1.57 microseconds, and an aggregate bandwidth of more than 61 Tb/sec.
InfiniBand QDR	The InfiniBand Trade Association (Industry Standard)	Quad data rate (QDR) InfiniBand has a signalling rate of 10 Gbit/s, which effectively provides 8 Gbit/s per link (implementers can choose between 1, 4, and 12 links). Switch chips have a latency of 100 ns, working with 4 KB messages.
Blue Gene/P IC	IBM	The PowerPC 450 chip integrates the logic for node-to-node communication, with a 5.1 GB/s bandwidth and 3.5 microsecond latency between nearest neighbours.
Tofu	Fujitsu	Made up of 10 links for inter-node connection with 10 GB/s per link, totalling 100 GB/s bandwidth organized in a 6D torus.

Table 7: Interconnect technologies

The most popular of these is InfiniBand QDR, used exclusively in seven machines and shared with InfiniBand DDR in Pleiades. Roadrunner is the sole machine to only use the slower InfiniBand DDR technology. Cray integrates their own interconnects in their machines, using the new Gemini interconnect in the XE6 Hopper, Cielo, and Hermit systems, and the older SeaStar 2+ in the XT5 machines (Jaguar and Kraken XT5). Both Tianhe-1A variants uses a NUDT custom designed proprietary high-speed interconnect called Arch that runs at 160

Gb/s. IBM uses its own proprietary interconnect for JUGENE, and K Computer has a custom built interconnect called Tofu, which is also used in the new Oakleaf-FX. Curie, on the other hand, sticks with the current market leader using InfiniBand QDR technology.

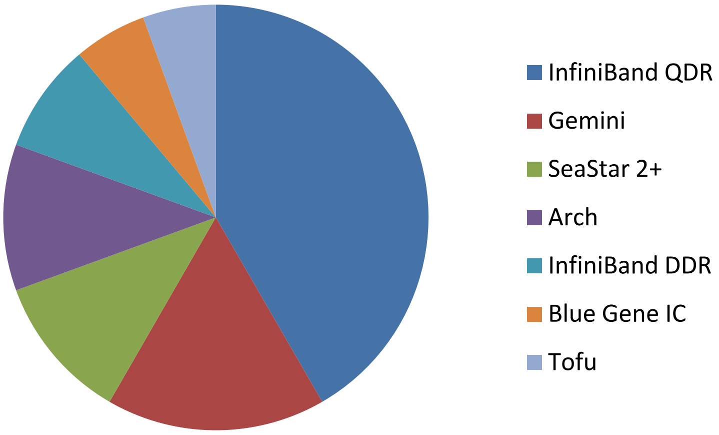


Figure 7: Interconnect technologies in petascale systems

Static analysis - Architecture

All petascale systems make use of multi-core CPUs, and all except one of them utilize homogeneous cores (meaning that all cores have the same architecture). The exception is Zin, which uses Intel's new Sandy Bridge heterogeneous processor that includes CPU and GPU on the same die. The only other heterogeneous processor employed is the PowerXCell 8i in Roadrunner and Lomonosov, but in these cases it works as an accelerator and not as a CPU. Multi-threaded CPU cores is a feature supported by Intel, Fujitsu, and IBM on their processors and avoided by AMD and ShenWei, giving an almost 60-40 percent split between the shares of both approaches. In the case of multi-threaded CPU cores, the number of threads per core is small (less than 8 threads per core). As was shown before, only 39% of the systems have off-chip accelerator units, and all of them are outside of the current standard programming models used on the main chip. NVIDIA Tesla accelerators use the CUDA programming model, while IBM PowerXCell 8i uses their own custom SDK. These specific languages are currently the most used, although OpenCL can be used for both accelerators. For more discussion on the different processor types and related programming models, [36] provides a series of annual snapshots with also some historical perspective.

The most common node configuration consists of heterogeneous nodes for different operational activities, which is used by half of the systems. Of the remaining machines, only four are heterogeneous at the compute node level (Jaguar, Tsubame 2.0, Pleiades and Lomonosov), while the other 5 are completely homogeneous.

The only industry standard interconnect used is InfiniBand (in its DDR and QDR configurations), and yet it still makes up for half of the interconnect types used. The rest of the systems use custom built interconnects designed by the vendor. In regards to topology, fat-tree and 3D-torus are used in about 40% of the systems each. The only other topologies used are higher-dimensional toruses: 4D in Pleiades and 6D in K Computer. For the new systems, Curie uses a full fat-tree network, and Oakleaf-FX the same 6D torus used in K Computer.

Memory addressing is generally uniform for intra-node memory. There is usually no global, intra-node, memory access. The only exception to this is the Cray XE6 system, which offers hardware support for Partitioned Global Address Space (PGAS), available in three systems (17%).

All of the petascale systems run some version of Linux. Specifically, Cray Linux Environment is present in four of them, and SUSE Linux Enterprise Server and Red Hat Enterprise Linux in two each. The rest use unnamed Linux distributions or specifically customized versions of Linux (Tianhe-1A and Sunway, for example). Tsubame 2.0 offers a dual operating system configuration with Microsoft Windows HPC and Linux. Curie and Oakleaf-FX follow the common norm and run varieties of Linux (RHEL in the case of Curie).

Handling of the parallel file system is mostly done through Oracle's Lustre (71% share), with Panasas PanFS in second place with a 12% market share. The other parallel filesystems used are IBM’s GPFS (JUGENE), Haskell’s HVFS (Nebulae), and a custom filesystem designed for the Sunway Blue Light. Both Curie and Oakleaf-FX use Lustre, which is taking a very strong lead in the market.

Static analysis - Infrastructure, packaging

Infrastructure resources and packaging options are possibly the most varying features in this analysis. Rack count is as low as 9 racks for the Sunway Blue Light system, and as high as 864 for K Computer (96 times more racks for little more than 10 times the peak performance). Floor space is directly related to rack count, and varies from only less than 50 square meters for Sunway Blue Light to 3,000 square meters for K Computer, which represents a more than 60-fold difference.

One of the most important details when studying petascale systems at the moment is power consumption. The most power-hungry machine on the list is K Computer, which consumes 12.66 MW while executing the LINPACK benchmark. In contrast, Mole-8.5 only consumes 0.54 MW. The average power consumption for all machines is around 3.3 MW, with most of them (11 of 18, or 61%) consuming less than the average.

At the moment, traditional air cooling methods are being complemented and replaced by liquid cooling solutions. Still, more than half of the systems on the list rely exclusively on air for their cooling (55%), but 28% are already using a heterogeneous air/liquid solution and 17% are cooling using liquid alone (Figure 8). Curie and Oakleaf-FX will both contribute to the disappearance of exclusive air-cooling, in Curie’s case in the form of a mixed air/liquid solution, as already does TERA 100.

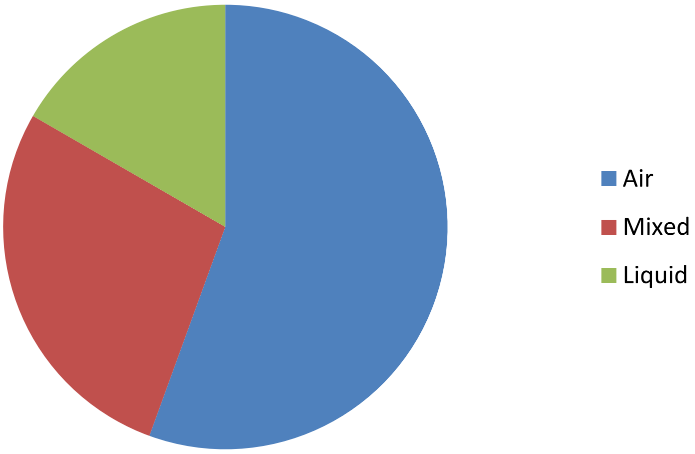


Figure 8: Cooling methods for petascale systems

Static analysis - Metrics

Many of the values we have analysed in the previous sections are not completely relevant by themselves, but need to be considered in relation to other values to be really interesting. For example, K Computer had the most racks, floorspace and energy consumption, but it is also by far the most powerful and also has the largest memory size. Therefore the ratio of these values to performance is much more useful.

Perhaps one of the most important metrics when analysing high-performance computers is the ratio between sustained and peak performance (Figure 9). Only three machines scored less than 50% on this ratio: Nebulae at 43% and Lomonosov and Mole-8.5 at 49%. The other three GPU-based computers (both Tianhe-1A solutions and Tsubame 2.0) perform between 52.1% and 57.5%. All other systems have a ratio that is between 74.4% (Sunway Blue Light) and 83.7% (Tera-100), except for K Computer, which breaks the mould with a stunning 93.2% performance ratio. The average performance ratio when including all systems is 72.53%, which grows to a little over 80% if the six GPU-based computers are omitted.

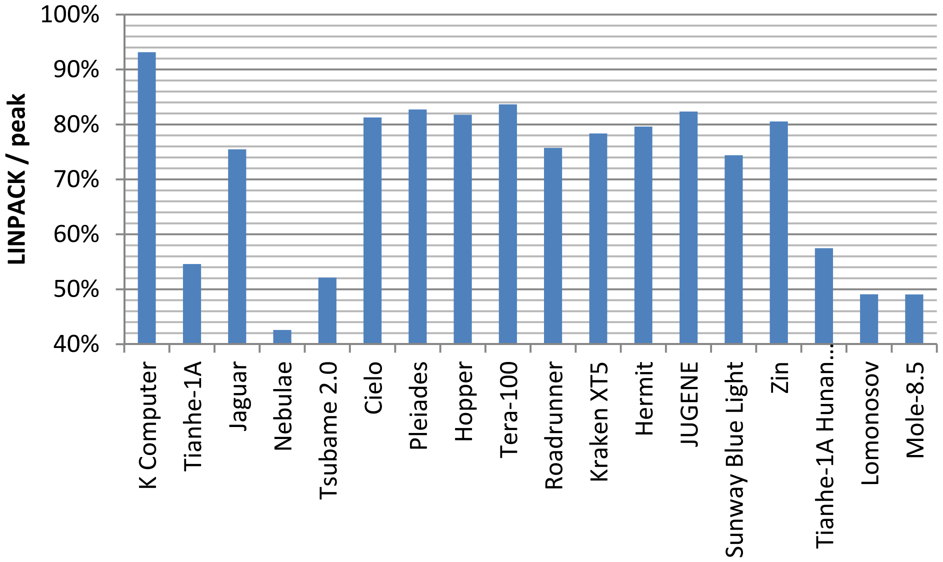


Figure 9: Sustained and peak performance ratio

Another very significant metric is the ratio between power consumption and sustained performance (Figure 10), which is used by the Green500 project to classify the most energy-efficient computers. Using this scale, Mole-8.5 clearly leads the others with its 920.4 Megaflop/s per Watt, 8% more than the next best result: Tsubame's 851.4 Megaflop/s per Watt. The least energy-efficient system is Tera-100, at 229 Megaflop/s per Watt. Traditional non-accelerated machines score quite low on this ratio (200-400 Megaflop/s per Watt), but new low-power processors like the RISC architectures Fujitsu VIIIfx and ShenWei SW1600 have brought non-accelerated systems to the same high-efficiency levels typical of accelerated systems (600-900 Megaflop/s per Watt). Going into more discussion would require longer developments, beyond the scope of this report. For instance Flop/s/Watt is an interesting metric but should probably be integrated in a broader vision of time/energy/cost to solution, which may vary according to a number of factors (the efficiency of applications relying itself on parameters like memory/core, interconnect bandwidth etc.).

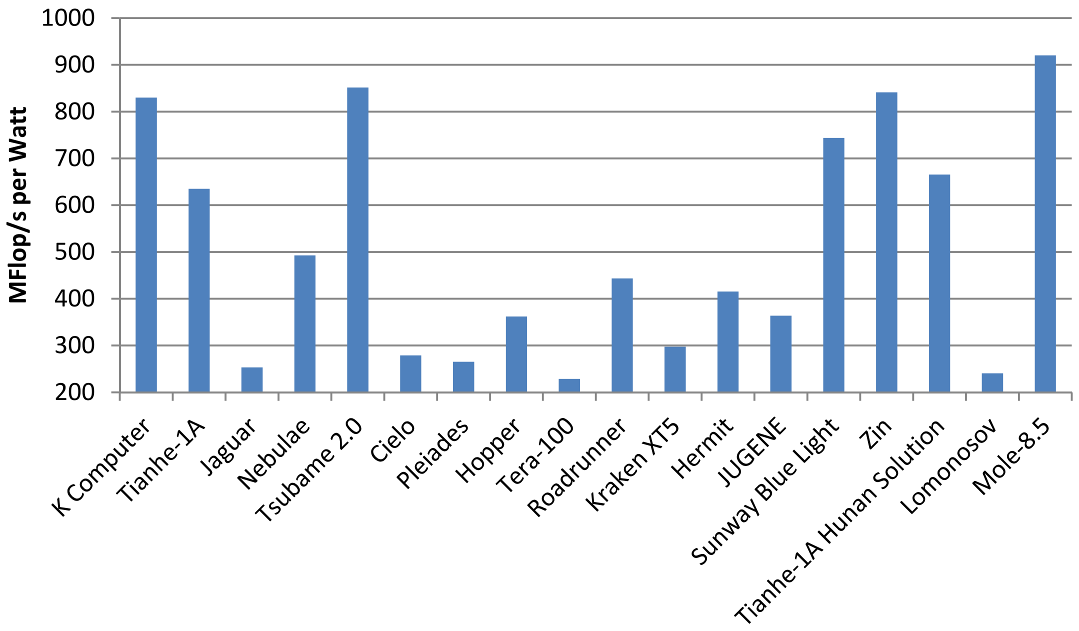


Figure 10: Power consumption and sustained performance ratio (MFlop/s per Watt)

With regards to memory distribution, an interesting metric is the amount of memory per node and core, which varies greatly depending on the system. In the per-node category, the lowest value corresponds to JUGENE, which has only 2 GB per node, and the largest amount is 71 GB, present in both Tsubame 2.0 and Tera-100 (71 is the average, Tera-100 nodes have 64 GB with some nodes having 128 GB of memory). The average memory per node is 30 GB, with several machines having a similar value (Tianhe-1A, Hopper, Roadrunner, Zin, and Cielo). Regarding per-core memory, JUGENE is again at the bottom with 0.5 GB per core, while Tsubame 2.0 leads at 5.8 GB per core. The average memory available to each core in petascale systems is around 2 GB.

The last metric that has been considered for this analysis is the number of cores per node, which gives a fairly good idea of the distribution of processing power. The lowest value of this ratio among petascale systems is 4 cores per node, which corresponds to Roadrunner and JUGENE. On the other side, Tera-100 has the maximum of 32 cores per node

2.2.2 *Dynamic analysis*

Having an overview of the current situation of the world-class HPC market is useful, but if one can compare this general view over time the interesting conclusions multiply. In this section we update this periodic overview, which was started in a previous PRACE Prepatory Phase deliverable [21], as well as in PRACE-1IP deliverables D8.1, D8.2, and D8.3 [26][27][28]. Understanding trends in supercomputing plans or roadmaps in different regions of the world is useful strategic information, in terms of sizing, timetable and manpower estimates for PRACE. A time-based analysis is made of several aspects of petascale supercomputers since their introduction, with an added note on future planned systems and their proximity to the observed trends. This section also gives a vision of petascale efforts based on a top10 analysis to position PRACE machines in "world-class" visibility. The data used in this analysis and for the generation of the graphs is available to PRACE partners in the project internal wiki.

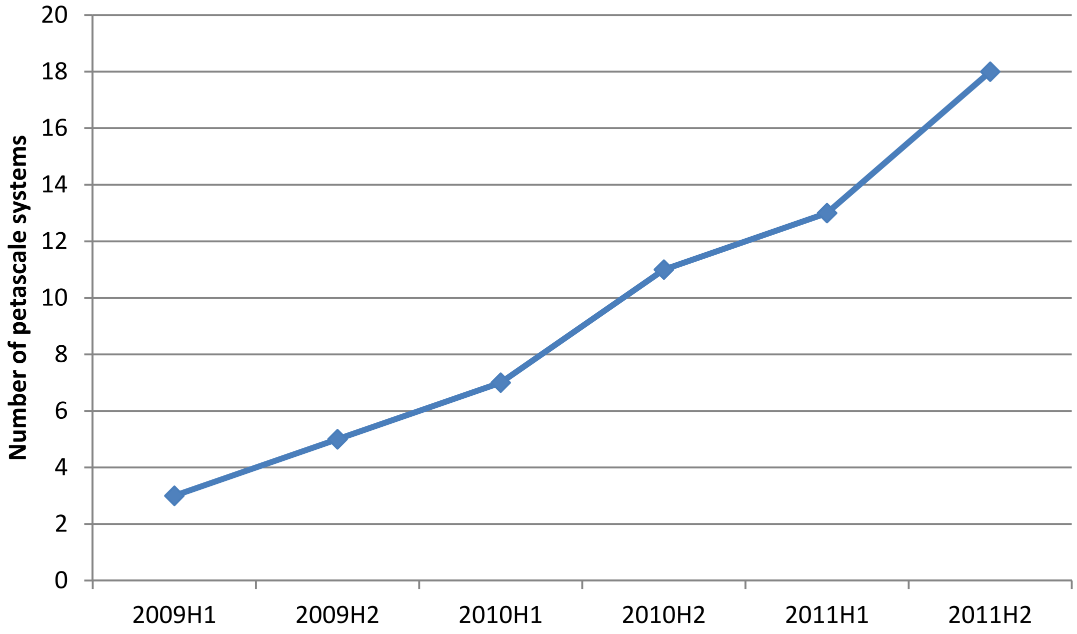


Figure 11: Increasing number of petascale systems (by half years / H1-H2)

Dynamic analysis - General information

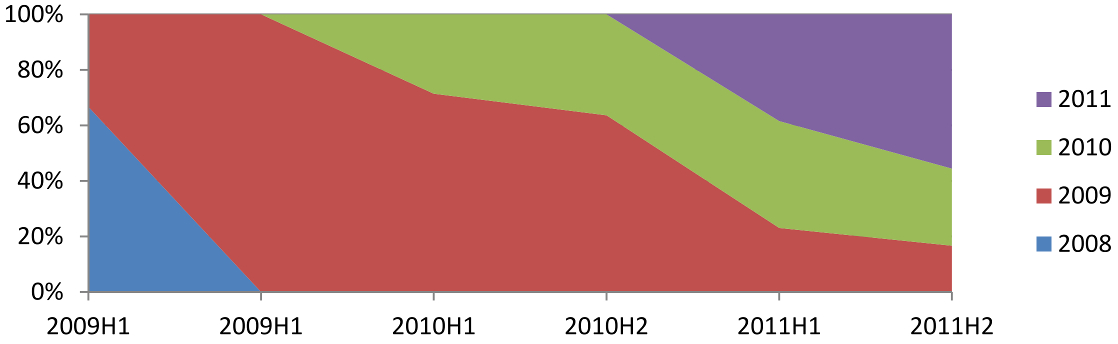


Figure 12: Evolution of the production year of petascale systems

Although petascale systems appeared in 2008, one year later they had all been upgraded and joined by new systems. The year 2009 represents the definitive entry of petascale machines as the prevailing top-class supercomputers (Figure 11). This tendency has steadily continued throughout 2010 and 2011, leaving the 2009 systems almost as historical relics (Figure 12). This will only be accentuated in the present year, when several of the older petascale computers are upgraded (most notably Pleiades and Jaguar). 2012 promises to be the year of petascale mass introduction, with the number of systems set to more than double current numbers.

Dynamic analysis - Country

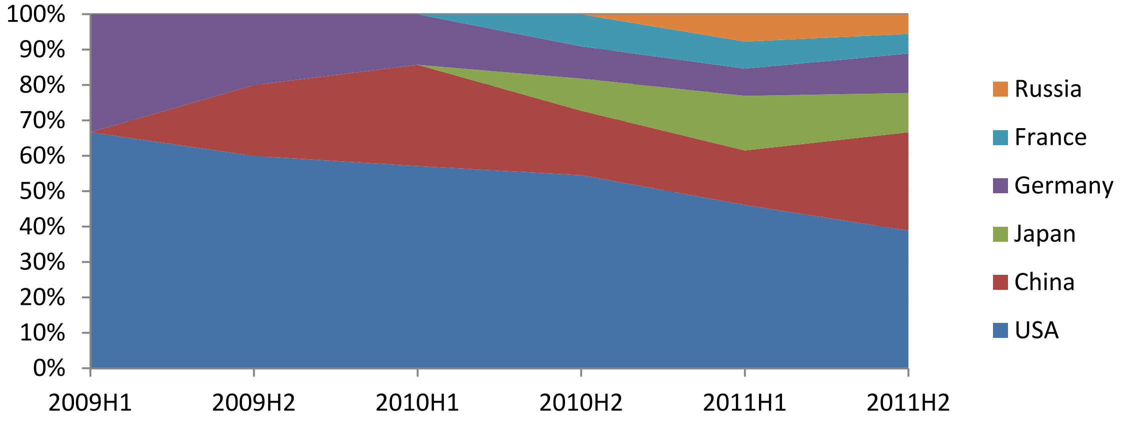


Figure 13: Petascale systems by country

It is not surprising that the USA has a dominant majority in petascale computers (Figure 13), but their share is continuously decreasing. The first half of 2009 shows the typical situation of earlier years: the USA almost exclusively commanding the share of top supercomputers, with Germany as the only other country present. In the second half of 2009 and start of 2010, China made an incredible jump from having no representation in the petascale market to controlling almost one third of it. The USA and Germany were unable to match this sudden growth by China, leading to a significant loss in their own shares. The entrance of Japan, France, and Russia into the competition in 2010 levelled the playing field, dropping both USA and China's market share considerably. Germany has made a slight comeback in the second half of 2011 by adding a second petascale system, but the most obvious advancement has been made by China, almost doubling their market share in one semester. Meanwhile, the USA is at an all-time low, maintaining their lead by a margin of only 11% share over China.

It is known that the USA has several petascale systems planned to enter production in 2012, but it is not clear whether China will have some of their own ready as well, perhaps even enough to definitively take the lead. For the time being, France and Japan will see their share grow slightly with their recent supercomputers, and the USA hopes for 2012 to be their comeback year, with six scheduled petascale systems (including first-place candidates Sequoia and Titan).

Dynamic analysis - Performance

Peak	**2009H1**	**2009H2**	**2010H1**	**2010H2**	**2011H1**	**2011H2**
Peak (PFlop/s)	1.46	2.33	2.98	4.70	8.77	11.28

Table 8: Evolution of peak performance of petascale systems

LINPACK	**2009H1**	**2009H2**	**2010H1**	**2010H2**	**2011H1**	**2011H2**
LINPACK (PFlop/s)	1.11	1.76	1.76	2.57	8.16	10.51

Table 9: Evolution of LINPACK performance of petascale systems

Table 8 and Table 9 show the evolution of the aggregated performance of petascale systems over years, resp. peak and LINPACK.

Dynamic analysis - Evolution & extrapolation vs. Announced projects

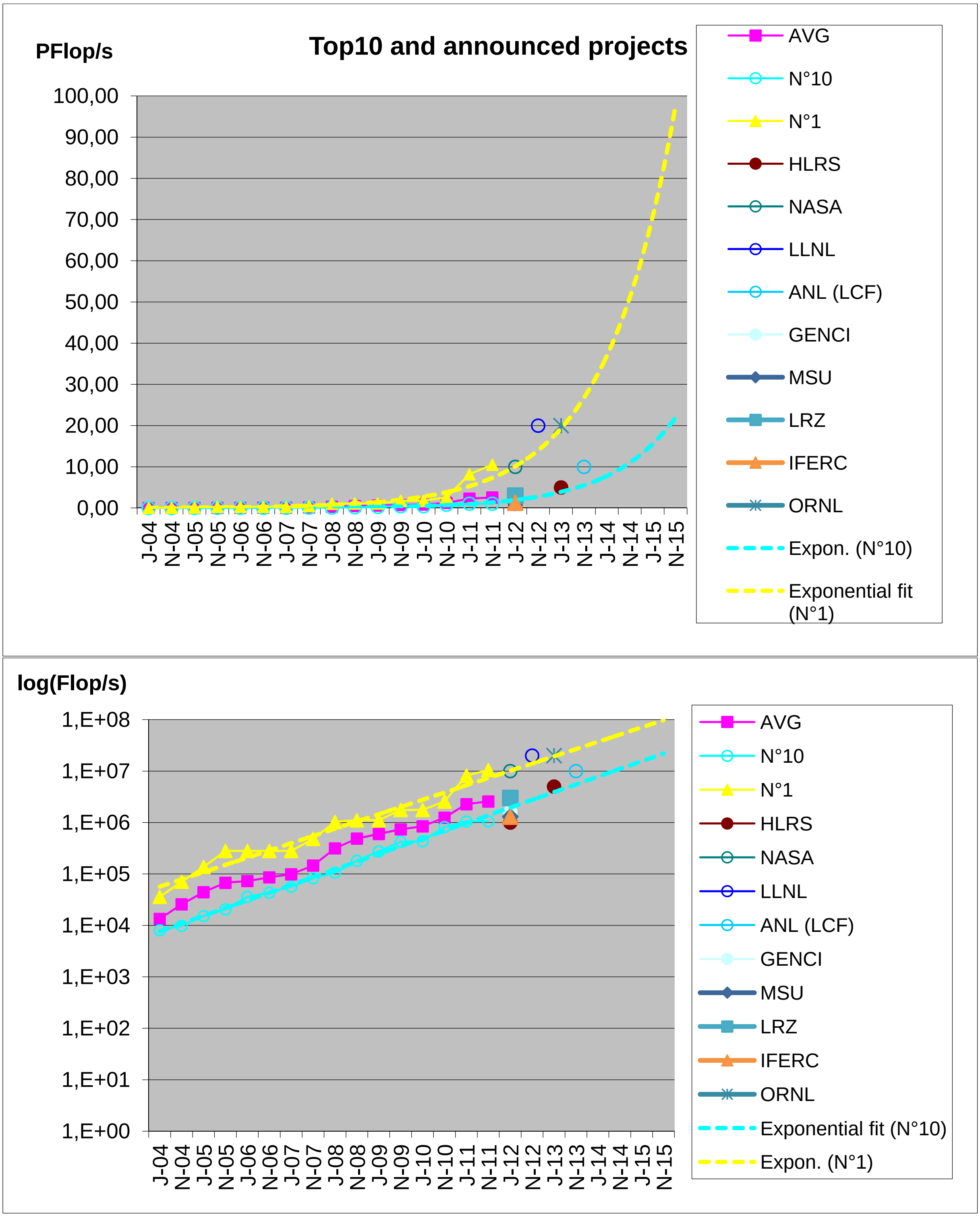


Figure 14: Top10 and announced projects

Figure 14 shows the peak performance of Top10 systems along time (each X-axis tick is a Top500 landmark, J=June, N=November), with some extrapolation to known projects, in both linear and logarithmic scales.

K Computer has been above the extrapolations presented in all previous deliverables, reaching the 10 Petaflop/s barrier almost an entire year before predicted. This large leap for Japan, similar to the one taken with Earth Simulator, is especially significant in the context of the earthquake that hit the country in March 2011, and in that the system is based on Japanese components (most importantly the Fujitsu processor). According to the information at hand, it isn't expected that the current leader will change until the release of either Sequoia or Titan (updated Jaguar), both scheduled to have a peak performance of around 20 Petaflop/s in 2012, but available towards the end of the year (acceptance testing on Sequoia is scheduled for September 2012). The rest of the high-end releases this year (Blue Waters, Carlsbad, and Mira) will fall slightly below or around K Computer's 11.28 Petaflop/s peak.

Predicting USA and Europe's future positions is relatively easy, considering public tenders mean early release announcements. In the case of China it is not so obvious. New systems are sometimes announced years in advance (the petascale project based on the Longsoon processor, which is still scheduled to arrive sometime this year), and at other occasionsdirectly upon inclusion in Top500 (as has occurred with the Sunway BlueLight MPP). There is no public release detailing any Chinese candidates for next month's Top500 list or the following November 2012 edition, other than the 1 Petaflop/s Dawning 6000. This does not mean that it shouldn't be expected that China will release competitive alternatives to the new American projects in the 10-20 Petaflop/s range, and perhaps some more based on Chinese processors.

Dynamic analysis - Vendors shares

	2009H1	2009H2	2010H1	2010H2	2011H1	2011H2
Cray	66.67%	40.00%	28.57%	36.36%	30.77%	27.78%
IBM	33.33%	40.00%	28.57%	18.18%	15.38%	11.11%
NUDT	0.00%	20.00%	14.29%	9.09%	7.69%	11.11%
SGI	0.00%	0.00%	14.29%	9.09%	7.69%	5.56%
Dawning	0.00%	0.00%	14.29%	9.09%	7.69%	5.56%
Bull	0.00%	0.00%	0.00%	9.09%	7.69%	5.56%
NEC/HP	0.00%	0.00%	0.00%	9.09%	7.69%	5.56%
Fujitsu	0.00%	0.00%	0.00%	0.00%	7.69%	5.56%
T-Platforms	0.00%	0.00%	0.00%	0.00%	7.69%	5.56%
Appro	0.00%	0.00%	0.00%	0.00%	0.00%	5.56%
NRCPCET	0.00%	0.00%	0.00%	0.00%	0.00%	5.56%
Tyan	0.00%	0.00%	0.00%	0.00%	0.00%	5.56%

Table 10: Petascale systems vendors across the years

A similar situation has occurred in the vendor market as in the country distribution: the historically dominant players (IBM and Cray in this case) have gone from an exclusive control of the petascale market at the start of 2009 to having to fight to maintain a controlling market share (Table 10). Luckily for them there isn't one strong vendor that is taking their entire market share, but a multitude of smaller vendors and institutions that are each pinching a small piece of the cake. Some of these will probably see their market share expand in the following years (Bull and Fujitsu will definitely grow with their new systems already in production, while SGI and Appro are also viable candidates for growth in the next years), while others will maintain their condition of experimental and/or non-commercial institutions (NUDT and NRCPCET).

Dynamic analysis - Technology bricks and components: processors

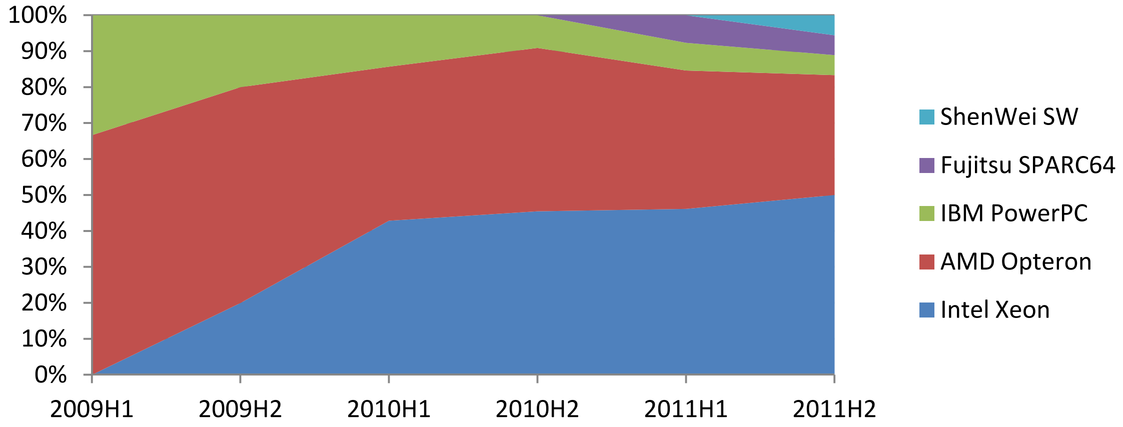


Figure 15: Processors used in petascale systems across the years

It is interesting to see in the distribution of processors that Intel, the overwhelmingly dominating manufacturer of processors for both consumer computers and high-performance supercomputers, was absent at the introduction of petascale systems and has had to catch up since then (Figure 15). In 2011 this had been accomplished and Intel was alone at the top of the market share list with exactly half of the petascale systems powered by their processors. AMD and IBM, which usually try to take a part of Intel's majority share, have in this case started with the dominant position and are striving to maintain as much as possible of it as Intel passes them by. The most surprising circumstance is the appearance of two other processor manufacturers in the list: Fujitsu and, more astonishingly, ShenWei. The Japanese and Chinese processor makers have ended the USA monopoly in this HPC segment, and may mark the beginning of a much more profound change in the processor market. It should be noted that these new processor lines are both RISC architectures (SPARC and DEC alpha inspired, respectively).

On the other hand, orders on IBM's Blue Gene/Q machines will contribute to higher market share of their IBM PowerPC processors. Either way, Intel and AMD seem to be losing ground in the top-class HPC market, and new players are entering at an astounding rate, including the Chinese ShenWei and other processors that are getting prepared to reach petascale (more in depth analysis on the HPC processor market can be found in the Business Analysis chapter).

Dynamic analysis - Technology bricks and components: accelerators

The introduction of accelerators paved the way for petascale computing, but hasn't yet consolidated a majority in the market. Similarly to what happened in the processor race, the first companies to introduce accelerators were not the commercial leaders (IBM and ATI, Figure 16), and NVIDIA has had to catch up to them with a fast growth. Now NVIDIA, which is the only manufacturer to commercialize specific HPC accelerator solutions, dominates this area. An interesting finding is that after accelerators made a strong entrance in the HPC arena, their share has been fairly stable throughout the past three years, floating around the 40% mark with very slight rises and drops. This has been mostly due to the introduction of low-power processors that achieve similar or better energy-efficiencies with much higher real performance due to hybrid programming issues. Curie and Oakleaf-FX illustrate this duality, with Curie using a hybrid architecture combining traditional general-purpose CPUs with HPC-specific GPUs, while Oakleaf-FX uses only one type of low-power general-purpose CPU.

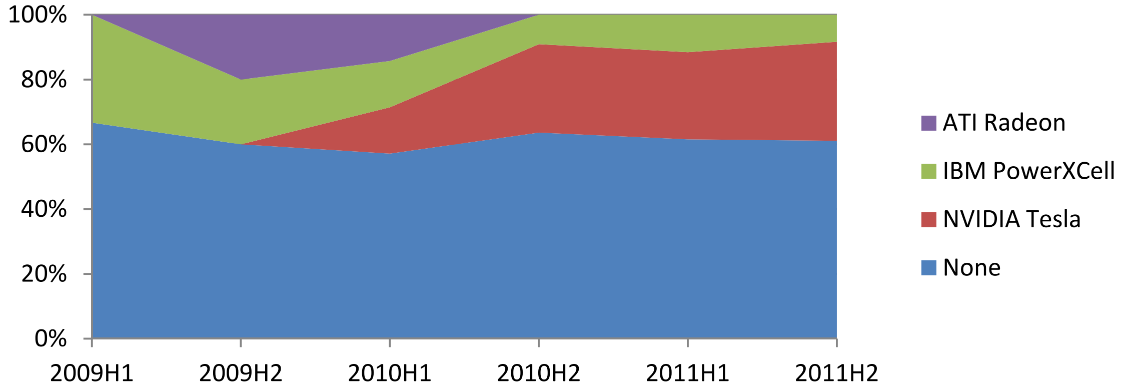


Figure 16: Accelerators used in petascale systems across the years

What seemed like an almost definite trend in the direction of accelerators has stabilized so far. The information available on upcoming petascale systems seems to point towards a decrease in accelerator use, although it also includes the introduction of a new kind of accelerator: Intel's Many Integrated Cores (MIC) architecture. If proven useful for HPC, it could mark the beginning of a resurgence of accelerators in supercomputing. On the other hand, perhaps ARM's introduction of 64-bit processors will challenge accelerators in their current form (more details in the Processor Market Analysis in the next chapter). IBM has abandoned their PowerXCell project, and ATI does not seem interested in commercializing specific HPC accelerator solutions, so that NVIDIA appears to have the upper hand in this derivation of HPC units from mass market accelerators. One important consideration is that Intel and AMD are both starting to roll out processors with incorporated accelerators, which may lead to heterogeneous processors replacing special-purpose accelerators.

Dynamic analysis - Technology bricks and components: cores

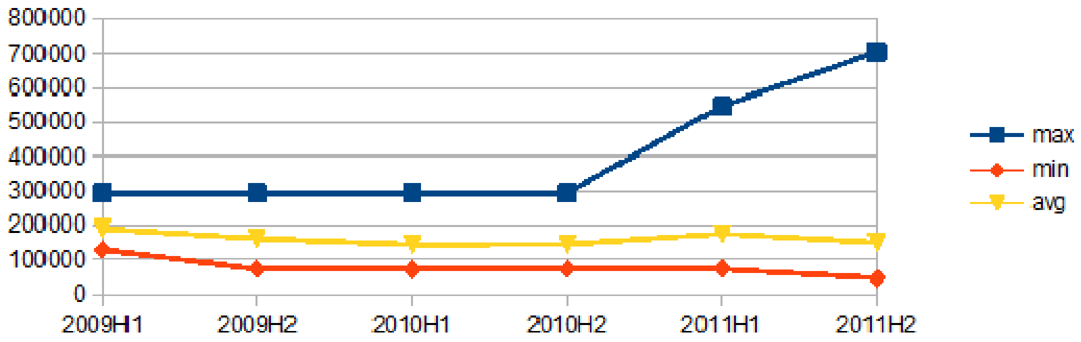


Figure 17: Number of cores used in petascale systems across the years

It is very hard to draw conclusions about the trends in core count of petascale systems since the maximum value is growing while the minimum is slightly decreasing. This is due to the two opposite routes used to achieve petascale: many low-power cores or accelerators. The average number of cores is more or less constant around the 150,000 mark (Figure 17).

Dynamic analysis - Technology bricks and components: interconnect

	2009H1	2009H2	2010H1	2010H2	2011H1	2011H2
InfiniBand QDR	0.00%	0.00%	25.00%	33.33%	34.62%	41.66%
Gemini	0.00%	0.00%	0.00%	16.67%	15.38%	16.67%
SeaStar 2+	33.33%	40.00%	25.00%	16.67%	15.38%	11.11%
Arch	0.00%	0.00%	0.00%	8.33%	7.69%	11.11%
InfiniBand DDR	33.33%	40.00%	37.50%	16.67%	11.54%	8.33%
Blue Gene IC	33.33%	20.00%	12.50%	8.33%	7.69%	5.56%
Tofu	0.00%	0.00%	0.00%	0.00%	7.69%	5.56%

Table 11: Interconnection types in petascale systems

Although it is still the most common solution, InfiniBand has had trouble maintaining its position in the petascale evolution. Due to the late introduction of their QDR variants, the market has looked to alternative technologies to achieve top-class performance (Table 11). Cray's proprietary solutions (SeaStar 2+ and Gemini) have taken a large portion of this free share, along with IBM's Blue Gene-specific interconnect, the Chinese experimental Arch and Fujitsu's Tofu solution. InfiniBand DDR is leaving the scene at an impressive rate, losing almost 80% of their market share in only 2 years. InfiniBand QDR use will probably continue to grow (definitely with Curie), and we will soon see InfiniBand FDR entering the market strongly (almost half of the upcoming systems analysed will be using this newer variant). Cray's Gemini interconnect is also expected to appear in several new petascale systems, and IBM will continue to use their proprietary interconnect in forthcoming Blue Gene machines (several Blue Gene/Q petascale systems are also expected in 2012). Fujitsu's Tofu interconnect is used in the new Oakleaf-FX system, and might appear in other new Fujitsu supercomputers.

2.2.3 *Business analysis*

General HPC market and trends

In our previous D8.2 deliverable, we reported from business intelligence and market survey sources [5][6][7][8][9] that HPC market was considered in good health with a resuming growth. Recent updates from similar sources [10][11][12] mostly confirm this. Storage and data management segments are the fastest growing ones, even if not the largest share of the market yet. Big data has become all the hype and is probably a promising growth segment for many providers.

Asia and Pacific region is clearly the most dynamic one in terms of growth rate, while EMEA has highest growth volume, and the USA market share tends to fall in percentage.

Everybody agrees on observing a wider accelerator adoption (mostly GPU, and among this, mostly NVIDIA but also Intel MIC which is showing up). Accelerated computing is a mix of a few very big (aforementioned) systems in the Top 10 and a growing number of smaller configurations, likely for testing and experimentation purpose in most cases.

Processors Market Analysis

At the moment we are in one of the most dynamic periods for the HPC processor market in decades, with new companies challenging the market leaders, as well as established companies making strong comebacks with their current offerings. Reaching the “power-wall” and looking towards the challenge of exascale computing has definitely provided the perfect breeding ground for innovations that may significantly change the future processor landscape. Another of the factors involved in this change is the enormous growth of the Chinese market, and their government's strong investment to try to make them the dominant country in HPC.

The following section will detail the current status of each of the main contenders in the processor market and their plans to expand their market share in the next few years.

1. **Intel**

Intel remains the market leader, and their newly released Sandy Bridge processor family has arrived with a slight delay but fulfilling performance expectations. With this new processor architecture Intel moves into the heterogeneous processor space (CPU+GPU in the same package), a path that is shared by AMD with their Fusion line. Intel has managed to make the transition to the 32 nm process, and is already making the transition to 22 nm with their upcoming Ivy Bridge processors starting mid-2012. This new update will also bring Intel's tri-gate transistor technology, which is expected to cut power consumption by 50%. They are also managing targets of 20% increase in CPU performance and up to 60% better integrated graphics performance, compared to Sandy Bridge. Achieving these levels of performance and energy efficiency will definitely help maintain Intel as market leader for the near future and maybe let them stay afloat above the rest of the market. On a different note, Intel is also entering the accelerator market with their MIC (Many Integrated Core) Knight's Corner product, which could end up giving them a leading position in this market, currently dominated by GPU vendors. The first petascale HPC system to use MIC accelerators is Stampede, scheduled to launch in 2013.

2. **AMD**

AMD is betting strongly on heterogeneous processors for the future, which became clear when they bought ATI (GPU manufacturer) in 2006, announcing their idea of a Fusion line of integrated CPU+GPU processors. Unfortunately for them, the process of combining both companies and their technologies was more complicated than expected, and the first Fusion products have only just recently been released, while Intel has already managed to assemble their own heterogeneous offering. In the CPU area AMD has just presented their new microprocessor architecture, codenamed Bulldozer, which will substitute the K10 microarchitecture. The most relevant change is the introduction of modules: a mix between a dual-core and a dual-threaded single-core processor. These modules are composed of two x86 out-of-order processing engines that share early pipeline stages (instruction fetch and decode), the FPUs, and the L2 cache. This supposedly improves performance-per-watt and reduces costs by freeing die area. It is perhaps early to conclude if this strategy will be successful, but independent reviews and benchmarks of the first desktop versions have already shown that perhaps expectations were too high. What has yet to be seen is Bulldozer's performance in HPC, which AMD promises will be dramatically better thanks to the new AVX extensions and fused multiply-add instructions. The first petascale HPC systems to use the new Bulldozer cores will enter production soon, possibly soon enough to appear on the next Top500 list, providing a better benchmark for the HPC industry. Bad news for AMD is that they will once again be behind Intel in the manufacturing process, since they aren't expecting to make the change to a 22 nm process until 2013 (with the so-called Next Generation Bulldozer). On a positive note, AMD's integrated graphics are still superior to Intel's offerings, at least until Ivy Bridge is released (and expectedly after), and their prices remain more competitive.

3. IBM

IBM is having a roller-coaster ride with their processor products, with POWER7 free-falling into possible extinction after stepping down from the initial Blue Waters project, while their PowerPC processors are skyrocketing thanks to the thriving Blue Gene line of supercomputers. Although the cancellation of the Blue Waters project doesn't imply the end of POWER7, as IBM is still marketing the chip for servers and even some HPC systems, it was definitely a strong blow for the company. They have already published some information on their upcoming POWER7+ and POWER8 processors, possibly trying to get Oracle's attention after they criticized Intel's Itanium roadmap. POWER7+, which should be released very soon (originally scheduled for Q3 2011), is a 32 nm shrink of POWER7, including higher frequencies, larger caches and on-chip accelerators. POWER8, scheduled for spring 2013, will be based on a 22 nm process and will have more cores, larger cache and accelerators. It is very hard to tell how these processors will perform, since IBM has not provided complete specific facts or targets.

What is becoming very apparent is that IBM's PowerPC line of processors used in Blue Gene are providing excellent performance-per-watt, with Blue Gene/Q systems taking up all the top positions in the Green500 list of energy-efficient supercomputers. Since energy-efficiency is becoming the top priority of funding agencies and clients in general, this could very well translate into a boom for IBM’s processor division.

4. SPARC

SPARC processors are experiencing a huge revival, with Fujitsu's SPARC64 VIIIfx powering K Computer's 11.28 Petaflop/s, reaching top place on the Top500 as well as establishing itself at the top positions of the Green500 list. This eight-core processor, introduced in June 2009, is manufactured using Fujitsu's 45 nm process technology. Almost immediately after the presentation of K Computer, Fujitsu announced the availability of their next processor update: the SPARC64 IXfx, which is part of their supercomputing product PRIMEHPC FX1, capable of reaching 23.2 Petaflop/s (currently the most powerful system with this chip is Oakleaf-FX, at 1.13 Petaflop/s). This update contains 16 cores (twice as much as the VIIIfx) at slightly lower clock-speeds (1.85 GHz), and provides 2 Gigaflop/s per watt, on par with the most energy-efficient IBM PowerPC chips.

Oracle has also recently introduced a SPARC processor called SPARC T4, designed to offer high multithreaded performance (8 threads per core, 8 cores per chip). This processor was originally in Sun Microsystems’ roadmap when the company was bought by Oracle Corporation, who has continued the development. Built at a 40 nm process size and running at frequencies between 2.85 and 3 GHz, it is the first Sun/Oracle SPARC chip to use out-of-order execution. Oracle has been criticizing Intel's processor roadmap for use in their servers, and perhaps the T4 represents their solution, or backup plan. In any case, it seems clear that Oracle doesn't plan on scrapping Sun's SPARC processor developments, and is investing in continuing this line.

Interestingly, Tianhe-1A has a number of nodes with SPARC processors (a total of 2,048), called FeiTeng-1000, which were developed in China based on OpenSPARC. This processor represents the third generation of YinHeFeiTeng (YHFT) series of processors, the first two being intended as stream accelerators or coprocessors specifically designed for HPC. It has already been announced by NUDT that the upgrading of the processor is under way, and that the new version will be used in future Chinese supercomputers.

5. ShenWei

ShenWei is the name of a line of microprocessors developed by Chinese company Jiāngnán Computing Research Lab since 2006, with intellectual property rights belonging exclusively to the People's Republic of China. Their latest (3rd generation) version, released in 2010, is a 16-core, 64-bit RISC chip operating at 1.1 GHz. Designated as SW-3 or SW1600, it is manufactured with a 65 nm process technology and has been used in the Sunway BlueLight MPP Supercomputer. The supercomputer uses 8,704 SW1600 processors to achieve a peak performance of 1.07 Petaflop/s consuming around 1 MW.

The technology used in these chips is still several years behind the previously mentioned companies, but taking into account they have reached this level in only 5 years it is not ridiculous to consider them a future player in the processor market. No announcement has been made regarding the roadmap for the next generation, but only a reduction in process technology to 45 or 32 nm (ShenWei skipped the 90 nm process and went straight from 130 nm in their 2nd generation to 65 nm in the third) could place them in direct competition with Intel and AMD (depending on pricing, of course). Certainly something to watch out for, if not because of direct competition in western markets, then for the possibility of controlling the Chinese and/or Asian markets and excluding the typical big players.

6. Loongson

Loongson is a family of Chinese general-purpose MIPS-compatible CPUs, fruit of a public-private partnership between the Institute of Computing Technology (ICT) and Jiangsu Zhongyi Group in 2002. Starting as an embedded 32-bit CPU (Loongson 1), it has advanced to become an 8-core 65 nm processor running at 1.05 GHz (Loongson 3B). Loongson doesn't support the x86 instruction set, but has hardware-assisted x86 emulation (achieving an average of 70% the performance). It is expected that in 2012 Loongson 3C will become available, with 16 cores using a 28 nm process and running at 1.5 GHz. If this is true, it might mean passing AMD or IBM in the process technology race and placing itself right behind Intel. The schedule might not be completely credible though, taking into account that from the beginning of 2011 there has been an announcement of a petascale supercomputer based on Loongon 3 processors, but the system has yet to be shown. Where Loongson chips have been seen is in the desktop and laptop market, where they are already starting to compete directly with Intel and AMD.

7. ARM

The ARM architecture, introduced in 1983, has been widely used since in all kinds of embedded systems and mobile devices. The news is that it is now making the transition to HPC, according to two recent announcements:

The European Union FP7 programme Mont-Blanc has started with the aim of creating an HPC system using low-power embedded technology provided by ARM. The initial prototypes installed at BSC (Barcelona Supercomputing Centre) use NVIDIA Tegra2 and Tegra3 processors, containing ARM Cortex-A9 cores implementing the ARMv7 instruction set. Although these prototypes will not be competing in the high area of the Top500, they will almost certainly take the top spot in the Green500 list, by a large margin. According to calculations, these new systems will use up to ten times less energy per Flop/s than current supercomputers. The real challenge will be adapting HPC codes to the ARM instruction set and CUDA, but the reward seems to be worth it.

The other important announcement is that ARM will soon be releasing the new ARMv8 architecture, including a 64-bit instruction set and extended virtual addressing for the first time. The new architecture supports all previous ARMv7 32-bit instructions, and will additionally support the new A64 instruction set. No implementations have been announced yet, although compilers have already been made available to certain partners, and operating system support is being worked on by Microsoft and the open-source community. It is still very early to tell how processors based on this new architecture will compete with traditional offerings by Intel and AMD.

ARM is also about to enter the accelerator market with a new line of low-power GPU accelerators called Mali.

2.3 PRACE in a global context

Analysing the worldwide top-class HPC scene provides objective conclusions on the evolution of the industry as a whole, which is very interesting in itself, but fails to transmit how PRACE is positioned in this global competition. This section will focus on describing and comparing PRACE supercomputers with respect to the rest of the petascale systems considered, regarding performance as well as hardware architecture and usage.

For this comparative analysis, the following six PRACE Tier-0 petascale systems are considered:

- JUGENE (GCS@FZJ, Germany) - 2009
- Hermit (GCS@HLRS, Germany) - 2011
- Curie (GENCI@CEA-TGCC, France) – 2012
- SuperMUC (GCS@LRZ, Germany) – 2012
- Fermi (CINECA, Italy) – 2012
- MareNostrum (BSC, Spain) – 2012
- JUQUEEN (GCS@FZJ, Germany) – 2012

2.3.1 *Computing power and performance*

The history of PRACE coincides perfectly with the advent of petascale, having the first PRACE Preparatory Phase Project started in 2008, the same year that Roadrunner became the first supercomputer in the world to achieve a sustained LINPACK performance of one Petaflop/s. Only one year later PRACE had its first Tier-0 system, JUGENE, at the petascale level, taking third place in the Top500 behind Roadrunner and Jaguar, both in USA.

While the rest of the PRACE Hosting Members prepared procurements for their future Tier-0 systems, China made it clear that they would be joining the USA and Europe in the petascale race by introducing Tianhe-1, which had a peak performance of over 1 Petaflop/s although it didn't manage to reach 600 Teraflop/s in LINPACK. Nebulae, introduced some months later, took second place in the June 2010 edition of Top500 with a peak performance of almost 3 Petaflop/s. By this time JUGENE maintained the number five spot on the list, while several other PRACE Hosting Member sites were already in construction of new facilities – or upgrade of existing ones - for their Tier-0 systems.

When Tianhe-1 was updated, in November 2010, it took first place in the Top500 list, reaching a peak of almost 5 Petaflop/s. By this time, the ten top supercomputers in the world all had a peak performance of at least 1 Petaflop/s, with JUGENE almost at the cutoff in ninth place. This list also saw the introduction of Japan in the petascale race, with TSUBAME 2.0 taking fourth place. At this point, the petascale market was dominated by the USA with five systems, while China and Europe each had two, and Japan one. While the French Tera-100 system, the second European petascale supercomputer after JUGENE, and the first one designed and integrated in Europe, is not a PRACE Tier-0 system, this list also included the first nodes of Curie, the second PRACE Tier-0 to enter production.

In 2011 Japan took the crown with K Computer, while PRACE presented Hermit, their next Tier-0 petascale system. By this time, the fifteen top systems in the world were petascale, with seven in the USA, three in both China and Europe (out of which two PRACE Tier-0 systems), and two in Japan.

As of today PRACE has four Tier-0 petascale systems, which will be confirmed in the next Top500 list in June 2012. JUGENE, Hermit, Curie, and SuperMUC have a combined peak performance of over 7 Petaflop/s, a significant fraction of which is reserved for PRACE. Later on this year, the two remaining Hosting Members will present their own Tier-0 petascale systems: Fermi in Italy and the new MareNostrum in Spain, adding another 3 Petaflop/s to make a total of 10 Petaflop/s available for PRACE researchers. Also expected before the end of the year is the substitution of JUGENE by JUQUEEN. How these systems will be placed on the list is very difficult to predict, since there are many other worldwide petascale systems projected for 2012, but it should be expected that all six systems occupy a place in the twenty-five first positions.

2.3.2 *Hardware architecture*

Although it is important to have standard metrics such as peak performance or LINPACK score to compare systems, it is equally meaningful to analyse how each supercomputer achieves those numbers by comparing their hardware architecture. This has become increasingly apparent with the use of accelerators in HPC, which provide impressive peak performance numbers that they can't manage to sustain in LINPACK or real-world applications. Although the basic cluster architecture (independent nodes connected through a high-speed network) hasn't changed, the need for improving energy efficiency has led to new interpretations. There are currently three “types” of clusters in the petascale race:

- The most traditional cluster, exemplified by Jaguar – and Hermit, Curie, SuperMUCin Europe, for instance - consists of high-power general-purpose processors, where each node is practically a server-class system. The main advantage of this architecture is that nodes are powerful enough to handle complex computations, reducing the stress on the interconnect. On the other hand, this architecture does little in the way of reducing power consumption.
- Hybrid clusters, made up of general-purpose CPUs connected with accelerators (typically based on GPUs), divide the computational burden between their two processor types. China's Tianhe-1A is a perfect example of a hybrid cluster. For certain data-parallel tasks, running on a GPU greatly increases performance, while reducing power consumption. The main drawback is that not all applications contain these optimal data-parallel tasks, and that even those that do must be modified to make this parallelism explicit (with programming models that are still in their infancy). Since it is also a new technology, communication between CPU and GPU and between nodes hasn't been fully optimized yet.
- With the introduction of multicore came the idea that increasing performance doesn't necessarily mean using more powerful processors, but simply use more of them. This has led to a new breed of supercomputers based on the integration of many low-power processor cores, as is the case of IBM's Blue Gene line, for example. This combination of low-power cores results in very high energy efficiencies, but also accentuates the need for a fast interconnect (since each core is only capable of relatively little computations) and highly scalable code (an application must be broken down into as many small pieces as possible to distribute it evenly across the system).

The first PRACE Tier-0 system, JUGENE, was also the first supercomputer to use a massive amount of low-power cores to achieve petascale performance. Since then, others have joined this movement (for example Fujitsu's K Computer) while IBM has upgraded their Blue Gene offering to its next generation, which will be used by the PRACE Tier-0 systems Fermi and JUQUEEN.

Hermit, on the other hand, is an example of the traditional high-power general-purpose processor cluster, using Cray's XE6 platform which combines AMD Opteron processors. Curie was set to also become an example of this class of cluster,however it had some accelerator nodes added in the summer of 2011 contributing an additional 200 Teraflop/s to the system. This doesn't seem enough to consider it a hybrid cluster (it only accounts for about 10% of the computer's total peak peformance), but does provide a testing bed for preparing applications.

While improving energy-efficiency can be tackled from a hardware architecture point of view (with accelerators and low-power processors), it can also be addressed through the cooling infrastructure. This is the case in both Hermit and Curie, which use a mixed air-water solution to improve their power consumption. The more recent SuperMUC represents the extreme of this method, utilizing direct hot water cooling on its Intel Xeon processors.

As can be seen from the preceding paragraphs, PRACE has a strong presence in both high-power and low-power processor clusters, but is missing the same level in the hybrid cluster segment. It cannot be said that one type of cluster is better or worse than the others, but having at least one of each is important so that different applications can target the architecture most fitting to its underlying algorithms. This is especially true for the case of hybrid architectures, since performance can see an order of magnitude improvement if the code is suitable.

2.3.3 *Usage*

In the race to be the top performer in supercomputing in the world, many times the usage of the machine is relegated to a second place, even though it is supposedly the most important aspect. Systems that have taken first place in the Top500 such as Roadrunner and Tianhe-1A have also been criticized for not serving a purpose above running LINPACK at a record-setting level. In this sense, PRACE is well aligned with their ultimate goal of providing researchers with the maximum amount of useful computational capacity. Where countries like the USA and China use supercomputing investment as a means of demonstrating their dominance, Europe tends to focus more on the added value for research.

The first Regular Call for PRACE saw 363 million core hours (all on JUGENE) awarded to 9 projects, and the numbers have only been growing since. The second Regular Call added 40 million core hours from Curie's fat nodes, while the third Regular Call doubled the original allocation and reached a total of 722 million core hours (with JUGENE, Curie and Hermit). The fourth Regular Call has awarded more than one thousand million core hours for European scientists, which is set to reach close to 1500 million core hours with the fifth Regular call – this latter one for allocations starting November 2012. According to these numbers, PRACE has been duplicating its allocated core hours every year.

Making comparisons at this level is fairly complicated, as no two projects are exactly alike, and many don’t even release information on usage (the case of most military-use supercomputers and in Chinese projects). The USA has several projects that are similar to PRACE, such as the former TeraGrid (now called XSEDE) and INCITE. The latter has announced core hours awarded at their Argonne and Oak Ridge Leadership Computing Facilities, with about 1600 million awarded to general scientific projects (while another 1100 went to discretionary use and for ASCR Leadership Computing Challenge), every year since 2010. It is expected that in 2013 this number will almost triple to 5000 million core hours granted. This shows that, while PRACE has been maintaining a steady growth from 363 to 1500 million core hours, INCITE has been still at a little over that amount for three years and is now expecting a strong leap. Even so, this only represents an approximate 1.5x improvement per year, lower than PRACE's 2x year-on-year gain.

3 Hardware-software correlation

This section aims to define different correlations and relations between architectures and applications. In terms of the PRACE-1IP project, this section represents an interface between WP7, dealing with petascale software, and WP8, which handles petascale hardware and infrastructure. Working towards this purpose of associating hardware and software at the petascale level it has become apparent that the initial goals set out in D8.1 and D8.2[26][27], i.e. to predict the performance of applications on any given hardware architecture and vice-versa, were unattainable for several reasons, mainly related to the impossibility of characterising and classifying both hardware architectures and software applications in a useful manner. It was therefore decided that a new approach would be taken, in which we would try to provide as much useful information as possible, conceding that it will never reach the point of providing a complete representation.

3.1 Programming Models

The most obvious interface between hardware and software are the programming models, which exist as an abstraction layer above hardware and memory architectures created to simplify programming. There are many different programming models according to the way in which they describe the underlying hardware. Although these models are not (theoretically) specific to a particular type of machine or memory architecture, almost all of them need some form of hardware support underneath. In this section we will analyse which programming models are supported by each of the petascale systems identified in the Market Watch.

The programming models have been classified into 5 groups according to the representation of the hardware architecture that they present to the programmer. These groups are:

- **Distributed Memory (DM)**: In these models each processor has its own memory, and nodes pool the memory of their different processors. Nodes are connected to other similar nodes through a network. Computational tasks can only operate on local data, and if remote data is required, the computational task must communicate with one or more remote processors. In general, programming models in this group are implementations of the Message Passing Interface, a standardized and portable message-passing system designed by a group of researchers from academia and industry to function on a wide variety of parallel computers. For this reason it is also commonly known as a Message Passing Model;
- **Shared Memory (SM)**: In this model tasks share a common memory address space, which they read and write asynchronously. Various mechanisms such as locks/semaphores may be used to control access to the shared memory. This model is usually used at the node level, allowing multiple threads and cores to access the same private processor memory;
- **Distributed Global Address Space (DGAS)**: These models have physically separate memories that can be addressed as one logically shared address space (same physical address on two processors refers to the same location in memory). Although this addressing simplifies programming, realizing it requires either hardware support or heavy software overhead. It also has the disadvantage of concealing data locality, leading to inefficiencies. For these reasons, and because newer models such as PGAS have achieved the same advantages without the drawbacks, the DGAS model has almost completely disappeared;

- **Partitioned Global Address Space (PGAS)**: This model assumes a global memory address space that is logically partitioned and a portion of it is local to each processor, simplifying programming while at the same time exposing data/thread locality to enhance performance. This model was developed very recently with the objective of reducing both execution time and development time;
- **Data Parallel**: These models focus on performing operations on a data set, which is regularly structured in an array, distributing the data across different parallel computing nodes. The most common example of data parallelism in HPC today is the use of Graphical Processing Units for general-purpose computing (known as GPGPU). Programming models have been created to help take advantage of the enormous parallelism of vector units in modern GPUs for tasks other than graphics processing.

The following Table 12 summarizes the programming models that are supported by each of the petascale systems presented in the Market Watch, organized according to the above classification.

	Distributed Memory	**Shared Memory**	**DGAS**	**PGAS**	**Data Parallel**
K Computer	Open MPI Fujitsu XPF	OpenMP	-	-	-
Tianhe-1A	MPI	OpenMP	-	-	CUDA OpenCL
Jaguar	Cray MPT	PThreads OpenMP SHMEM	-	UPC CAF	-
Nebulae	<no info>	<no info>	-	-	CUDA OpenCL
Tsubame 2.0	Open MPI MVAPICH2 MPICH2	OpenMP	-	-	CUDA
Cielo	MPICH2 SHMEM	PThreads OpenMP	-	UPC CAF	-
Pleiades	SGI MPT MVAPICH2 Intel MPI	OpenMP	-	-	-
Hopper	Cray MPT MPICH2 SHMEM	PThreads OpenMP	-	UPC CAF	-
Tera-100	BullXMPI IntelMPI MPC*	PThreads OpenMP MPC*	-	-	CUDA HMPP
Roadrunner	Open MPI	PThreads OpenMP	-	-	IBM ALF LANL CML
Kraken XT5	Cray MPT	PThreads OpenMP	-	UPC	-
JUGENE	MPICH2	OpenMP SMPSs	-	-	-
Lomonosov	<no info>	OpenMP	-	-	CUDA OpenCL

Table 12: Programming models vs. petascale systems

** MPC is a framework that implements several programming models with a number of compatibility and performance enhancements (PThreads, MPI, and OpenMP).*

By looking at the table we can see there aren't too many different options in the programming model field. In each category there is a maximum of about 4 possibilities, and this doesn’t take into account that most of the alternatives in the Distributed Memory group are variations of the same Message Passing Interface (almost all based on the same OpenMPI or MPICH implementations). Distributed Memory and Shared Memory are by far the most popular groups, present in every petascale machine. On the other hand, DGAS is not available on any of the systems, having been somewhat superseded by the PGAS technology, which tries to combine the simplicity of programming DGAS without losing data locality, and minimizing hardware and software overhead. Even so, the PGAS model is only available on four of the petascale systems, two Cray XT5 systems and two Cray XE6 systems (with hardware support for PGAS through their Gemini interconnect).

3.2 Benchmarks and applications

In this section we will consider the possible techniques which can be used to understand how different applications will perform on different hardware systems.

To begin with, we note that it may not be very helpful to use architectural classes for this work: in the current era it is difficult to define classes of architecture which are not so broad as to be almost useless, or where the differences between systems within a class are as significant as the differences across classes. Notions such as MPP, thin-node cluster and fat-node cluster which have been used in previous PRACE deliverables now feel somewhat out-dated. It is probably of more use to consider how well applications might run on specific systems, rather than to try to characterize their suitability for broad architectural classes. A recent analysis of the Top500 data used just three classes: “heavyweight” (high capability power-hungry cores), “lightweight” (lower capability, power efficient cores) and “accelerated”. According to this classification, the current petascale systems are divided as follows:

- **Heavyweight** Jaguar, Helios, Cielo, Pleiades, Hopper, Tera-100, Kraken, and Hermit
- **Lightweight** K Computer, Oakleaf-FX, Sunway Blue Light, and JUGENE
- **Accelerated** Tianhe-1A, Nebulae, Tsubame 2.0, Curie, Roadrunner, Lomonosov, Tianhe-1A Hunan Solution, and Mole-8.5

There are a number of different approaches, which can be taken to try to measure or predict the performance of an application on a given system. Of course, benchmarking is the most accurate, but may not be feasible if access to a suitably sized system is not possible. Full simulation is a possible approach, but requires a very significant investment in tools to support it. It is computationally demanding for large applications and machines and requires a high degree of expertise.

A combination of performance modelling and limited simulation (of, say, the interconnect but not of the individual nodes) is an approach which is more tractable, if less accurate than full simulation. If we can construct performance models for key applications (which is underway in Task 7.4 of PRACE-1IP), then it is possible (with some constraints) to answer the question "How well will this application perform on an architecture which does not exist, or we do not have access to, but have a description of in terms of a fairly small set of parameters?".

Since a useful metric on the space of applications has yet to be demonstrated, it is not possible to conclude that, if Application A is "close to" Application B, then Application A will perform in a similar way to Application B on Architecture X. The sorts of metric which are sometimes used (e.g., Application A is in the same science domain as Application B, or

Application A uses the same broad class of algorithm, e.g., 7 dwarves, as Application B) seem to have little if any predictive power. Therefore, any performance models and simulations would have to be specific not only to the hardware architecture, but also the software application.

A less formal approach, which can make qualitative but not quantitative predictions, is to measure characteristics of an application which can then be used to understand how the application might be sensitive to architectural parameters. Interesting characteristics are:

- Percentage of instructions which are floating point;
- Number of Flop per load/store;
- Ratio of stores to loads;
- Number of cycles per L2 /L3 /main memory cache reference;
- L1, L2 and L3 cache miss rates (as percentages);
- Number of cycles per byte communicated;
- Percentage of time spent in each communication routine used;
- Distribution of message sizes in each communication routine used;
- Percentage of time spent in I/O operations;
- Distribution of data sizes in I/O operations;
- Memory utilisation (e.g., min/mean/max high watermark across MPI tasks).

These metrics can be measured (or derived) quite straightforwardly by using hardware counter instrumentation and communication profiling, as demonstrated in PRACE-PP Deliverable 6.2.2. Some of these metrics are not exclusively dependent on the general architecture, as they may depend on parameters such as the sizes of caches, but if we measure all the applications on the same machine comparisons can be useful. Memory bandwidth would be useful to measure (note that it is in general not the same as 1/[Number of cycles per main memory reference] due to hardware prefetching), but unfortunately is not easy to measure in practice.

The ability to collect data such as this raises the question as to whether predicting performance using a statistical analysis of these metrics on different architectures might be feasible. One can also characterize architectures using parameters such as peak Flop/s rate, cache sizes, memory latency, memory bandwidth, network latency and network bandwidth. Theoretically, statistical methods could potentially be used to predict the performance of a given application on architectures with given parameters. However, the number of parameters involved (for both applications and architectures) would require a very large number of measurements of different applications on different architectures, and it is not clear whether useful predictive power can be obtained. This is probably therefore not feasible in the context of the current project.

3.3 User survey results

As part of WP7, a survey was carried out among the users of the two first Tier-0 systems in production: JUGENE and Curie. Included in that survey were several questions, proposed by WP8, related to hardware use and requirements. The results of the survey provide a more indirect (though arguably more useful) idea of the relationship between hardware and software in the petascale era. The questions, along with the conclusions extracted from the responses, are as follows:

- What is the name of the application code used in your PRACE project?

Out of 54 responses to this question, 40 different applications were mentioned. This proves what has been stated in the previous section: applications are diverse and heterogeneous, and therefore difficult to classify in a meaningful way for our purposes.

- Which parallelization method does your application use?

There is an almost 50-50 split between two parallelization methods: MPI (which representes an exclusively Distributed Memory approach) and MPI+OpenMP (a hybrid solution with both Distributed and Shared Memory), with the first having a slightly higher share but on a steady decline compared with previous surveys. The answer to this question is fairly biased since the two Tier-0 systems on which the survey is based support only these two models, although Curie will support Data Parallel models with their new GPU accelerated nodes. The third production Tier-0 system, Hermit, was not included in this survey because it hadn't been used by PRACE users when the survey was launched, but in future versions it could provide a fourth programming model: PGAS.

Although it is true that MPI and MPI+OpenMP are the most popular parallelization methods in current HPC applications, their exclusivity in the results of this user survey is misleading. This question will probably bring much more useful information in future surveys, when Curie's GPU nodes and Hermit are included.

- What memory size per core is required for your typical production jobs?

The replies to this question were:

- 70% less than 1 GB
- 15% 1-2 GB
- 9% more than 2 GB
- 6% not specified

Again, the results must be taken with a grain of salt since JUGENE only provides 0.5 GB per core. When more Tier-0 systems become available the results will probably even out and be more meaningful. In any case, what can be extracted from this data is that application requirements can, to a certain degree, be modelled according to the hardware constraints. It will be interesting to see how users adapt applications to new Tier-0 machines with different architectures.

- What is the minimum amount of disk space required per production job?

Answers vary widely in this case, with the 100GB–1TB range slightly dominating with 32% of the responses, the upper sector (>1TB) following with 29%, and the 10GB-100GB range with 18%. It seems clear that storage requirements from a user's perspective can be summarized as "the more the better".

- To try and assess the requirements for the PRACE systems quantitatively, we would like you to score the following architecture features in terms of importance to your code. A total of 20 points should be distributed amongst the following requirements, with higher priority requirements receiving a higher number of points.

The reliability of the results for this question depends on the user's understanding of the underlying bottlenecks in their application execution, but it isn't difficult to predict the highest scoring answer: "higher peak Flop/s rate". For general-purpose CPUs, peak Flop/s rate is almost directly proportional to application performance, and therefore the most sought-after feature for users. Until recently, maximizing peak performance was fairly simple, by moving to smaller lithographic processes and increasing clock-rate of the CPU. Nowadays increasing peak performance is not so easy and, depending on how it is achieved, doesn't necessarily improve all applications' performance.

The second most requested hardware feature is "lower point-to-point communications latency", followed closely by "Higher memory bandwidth". These have been identified for some time as the most important bottlenecks in supercomputer design. Interconnect technology advances have been a major contributor in the petascale race, with many new players trying different strategies to reduce the bottleneck, like the Chinese Arch, Fujitsu's Tofu, or Cray's Gemini. Memory bandwidth, on the other hand, is not advancing at a similar rate. Memory technology has been relatively stagnant, and although capacity increases exponentially, latency and bandwidth are not improving at a similar rate. From a hardware point of view, this won't change dramatically until the introduction of new memory technologies (3D memory, MRAM, T-RAM, Z-RAM, etc.) which are not expected very soon. Until then, the best strategy for improvement is optimizing cache levels and cache management through both hardware and low-level software.

The next most needed features are "lower memory latency", which is closely related to what has already been mentioned for the "higher memory bandwidth". The least important features for users are "higher I/O bandwidth" and "higher bisection communications bandwidth", showing that optimizing interconnect latency is much more meaningful to users than improving the bandwidth (especially with the new breed of supercomputers based on enormous amounts of low-power cores).

- Are there any other architectural features that might affect the performance of your application?

Only 15% of users answered this question, with responses including "higher collective communication performance" and "larger NUMA regions".

- Could your application benefit from accelerator devices, such as GPGPUs?

This is a very interesting question, since the use of accelerators is known to improve peak performance dramatically, but improvements in real application performance are much lower (or inexistent in some cases). A total of 61% of the responses indicated the applications have accelerator implementations (20% "application has already been ported to accelerators") or may potentially benefit from accelerators (41% "application has potential to exploit accelerators but port has not been done or is in progress"). Of the remaining 39% of responses, 21% were confident that their application was unlikely to benefit from accelerators, while 18% were not sure. In any case, it seems that there is a majority of users interested in accelerators for their applications, and hopeful that it will improve performance.

3.4 Conclusion

Programming model support in petascale systems has proven to be a good way of laying the foundations for a hardware-software correlation, in that it represents the interface by which programmers are able to visualize and interact with different hardware architectures. It obviously doesn't provide any knowledge about the performance of the systems or the programming models, but does help to classify them in some way, as well as establishing a preliminary interface between the hardware and software layers.

On the other hand, performance data can be partially extracted by combining several methods (essentially performance models, simulations, and benchmarks) as well as analysing individual application parameters, but the requirements for these predictions to be useful (in terms of the number of software and hardware specifications necessary) make them unattainable in the near future. This will probably change, since performance models are relatively new and progress is accelerating.

Although not an objective empirical analysis, results from user surveys do help in providing indirect information on application and hardware requirements. The current user survey is based on a very limited number of systems and users, and therefore the results are not completely meaningful, but the method does seem productive.

As more Tier-0 machines enter production PRACE will have a better overview of application performance across different architectures. By studying the underlying programming models, running benchmarks, and receiving input from the users, useful information can be gathered to further advance in the matching between hardware architectures and software applications.

4 Decision making for planning and upgrading HPC centres

4.1 Objective and approach

Experience in the consortium has shown that the people tasked with the planning of new HPC centres and upgrading existing ones are often not construction or facilities specialists and thus have to acquire a lot of knowledge on the job whilst sticking to a very stringent timeline.

Task 8.2 of WP8 aimed to collect the combined experience of the consortium partners in planning and operating supercomputing centres in order to produce a series of white papers on key topics related to infrastructure planning. These papers aim to support those responsible for future infrastructure projects in their decision-making. This way the learning curve of each individual can be shared and made valuable to the whole consortium.

The output does aim to provide a recipe for how to plan and build a supercomputing centre as this will differ substantially from case to case depending on a vast number of factors.

For these studies, a number of important topics that decision makers should take into consideration during an infrastructure project were defined. These topics were each developed separately in form of white papers, thus producing an information "package" that decision makers can use.

The lead for the various papers was alternately taken by Norbert Meyer (PSNC), Erhan Yilmaz (UYBHM) and Ladina Gilly (CSCS). Information for these papers was sourced from the entire PRACE community and especially from contributors to this work package and task.

The white papers mentioned in this section are being made publicly available on PRACE web site (http://www.prace-ri.eu/whitepapers). The remainder of this section merely extracts executive summaries and recommendations so as to serve as quick reference and to give a synthetic view of the main findings and outcome of this effort.

Survey questions associated to the different topics and which fed the white papers are given in Annex 8.1 to 8.5.

4.2 Abstracts and recommendations from the white papers

In this section we only quote excerpts of the white papers, namely abstracts and recommendations (each white paper has such a recommendations section). For more background and self-contained explanations, please refer to the full white papers as referenced above.

4.2.1 *Selecting a Location for an HPC Data Centre*

Abstract

Choosing the location is an important strategic task when planning a new HPC centre, as it will impact virtually every step of the planning and realisation process as well as the future operation of the centre and extension possibilities. It is not uncommon in the data centre industry to spend significant effort and resources on finding and acquiring the optimal site for a new data centre. It is therefore unsurprising, that the industry has written widely about the factors that ought to be considered, when selecting a future location. However, within the community of HPC research centres the requirements in a number of areas tend to vary from those of the traditional data centres.

Based on a survey of the PRACE community to which 10 sites responded this paper sets out to discuss where requirements and options in terms of the search for a new location for an HPC centre differ from those of a traditional data centre and briefly discusses the criteria that this community attributes the most importance to when selecting a site.

Recommendations - Important criteria for selecting a new location for an HPC centre

Overall, the accumulated experience of PRACE sites shows, that the most important criteria for selecting a new location for an HPC data centre are the following:

- Size, shape and geological quality of the land
- Availability of sufficient and extendable power supply
- Availability of good communication network
- Availability of sufficient space for future extensions of the data centre
- Possibilities for use of free-cooling
- Proximity of R&D environment
- Easy access for deliveries and visitors
- Limitations due to local building, fire and security regulations
- Limitations due to the protection of local fauna, flora, water etc.

4.2.2 *Cooling – making efficient choices*

Abstract

Due to the increasing role of energy costs and growing heat density of servers, cooling issues are becoming very challenging and very important. While it is commonly accepted that power consumption is the number one challenge for future HPC systems, people often focus on the power consumed by the compute hardware only, often leaving aside the necessary increase in cooling power, which is required for more densely packaged, highly integrated hardware. The information presented herein is a result of data analysed and collected in a process of distributing a detailed survey among PRACE partners. In the paper we go into particulars of the cooling area by presenting different technologies and devices currently used in modern HPC data centres. We also try to describe innovative and promising solutions adopted by some PRACE partners that may pave the way for future standards. We focus on highlighting all advantages and disadvantages of each described solution. In the final part we try to provide general recommendations for HPC centres required to be taken into account when building an appropriate cooling system.

Recommendations

Based on the collective experience of partners from different partner organisations, spanning both current operations of HPC centres and insight from the challenges posed by future supercomputing systems, here is a list of main recommendations that should be taken into consideration during the process of designing a new infrastructure or upgrading an existing one:

- Try to use direct water cooling for the high density (more than 10 kW/rack) systems.
- Include additional water loop pipes for hot water cooling in your facility whenever you are designing a new HPC centre
- If forced by external conditions (old building etc.) to use air cooling, use hot and cold air separation techniques
- Try to exploit benefits of the location of your Data Centre: low ambient air temperature, lake or river in close vicinity etc.

- When using direct water cooling try to employ a system that allows for operating at temperatures above 40 degrees Celsius. This will simplify your cooling system, make it more efficient, reliable, and allow for simpler heat re-use.
- Currently there is no single good solution for all IT equipment: flexibility of air cooling comes at the price of efficiency. Consider what is the most efficient choice for each class of your equipment.
- When buying new systems, consider asking for the PUE of the machine (including chilled water generators etc.) to be at most 1.1.
- Whenever possible, use TCO metrics, including at least costs of power and cooling for predicted period of use of the new building or supercomputer, instead of mere acquisition and maintenance cost.

4.2.3 *Electric power distributionin HPC Centres*

Abstract

The design of the electric power distribution network for an HPC centre is an important strategic task when planning a new centre or planning the upgrade of an existing one. All decisions taken at the design stage may have long-term effects on the operation, maintenance and later upgrade of the power supply infrastructure.

Based on a survey of the PRACE community, this paper describes common issues related to HPC centres power distribution, discusses available solutions to the common problems associated with electrical energy distribution and eventually gives recommendations that may be useful for general or technical managers facing a new facility design or upgrade project.

Recommendations

This chapter presents a set of recommendations and good practices, based on the experience of PRACE partners, analysis of the current state of art and practices of designing big facilities. The goal was to support Tier-0 and Tier-1 sites at the process of upgrading or designing new electric power facilities.

Recommendation 1: Provide only the necessary level of redundancy

- Understand the expectation of the users in terms of required availability of systems. Carefully consider what level of redundancy is required for which systems and what the percentage of the total power load is.
- Redundancy should only be implemented where required since it has impact both on investment costs and on running costs.
- Make sure to get statistics about the quality of electricity delivery at your site for the past years in order to take a well-thought-out decision. Good quality of the electricity source by which the centre will be powered may allow decreasing the redundancy level and further maintenance costs.

Recommendation 2: Implement redundancy in a ‘clever’ way

- Provide at least two independent main supplies, sized to match calculated maximum demand of power. Additionally, the supplies can be further sized so that one of them covers with its capacity most of the connected load, whereas the second one is sized mainly to take care of the critical loads/systems. It practically means that the second system is much smaller than the main UPS.
- Optimise the mix of generators, UPS, distributed short-term UPS devices.

Recommendation 3: Internal experts

- Have in-house a team of infrastructure experts or at least qualified staff for infrastructure issues
- Involve this team as much as possible in the decision making process about future systems.
- Visit HPC centres of a size similar to yours and attend conferences on infrastructure issues.

Recommendation 4: Think TCO

- Power distribution equipment has typically a very long life time. This should be taken into account when comparing TCO costs.

Recommendation 5: Reduce electricity conversion losses

- It is desirable that high voltage, from a technical point of view, should be as close to the load as possible. For small and medium-sized computing or data centres it is clear that medium voltage supply is the most efficient. For Petaflop/s and Exaflop/s facilities it may be more economical to request higher supply voltage, usually in the range between 100 and 137 kV.
- Reduce costs of power losses in a power distribution network through the use of low-loss distribution equipment or exchanging high-loss devices into new technology systems.
- Take into account the real efficiency of power distribution equipment under the expected load.

Recommendation 6: Introduce a design phase in order to make future evolution easier

- Design a network using modular approaches: i.e. one transformer per Low Voltage(LV) switchboard section, two interleaved downstream UPS units per LV section. This will certainly help to keep the infrastructure design coherent and will decrease design errors or inconsistencies.
- Always ensure that the distribution system is designed in order to account for future expansion. 30 to 35 percent of spare capacity on both the electrical supply size as well as on physical size of distribution equipment is a typical over-commitment.
- The electrical network must be "future proof" and the process of selecting equipment should ensure that the infrastructure provider will guarantee to have spare and replacement parts available in the future.
- If installation of water based cooling systems is considered, try to separate the placement of electric busbars and water pipes, for example, place the busbars above the racks and water pipes under cabinets (under the raised floor).

Recommendation 7: Monitor the power distribution system

- Deploy a monitoring system throughout the distribution network in order to ensure more efficient use of the available energy and facilitate the maintenance of the network.
- Assessment of the condition of the power distribution network should be carried out annually and after any modification to the existing electrical infrastructure. The results of technical condition assessment together with the data from the monitoring system should aid in planning any upgrade and maintenance works.
- For the technical condition assessment the use of modern diagnostic methods such as thermal imaging and vibration monitoring devices is advised. They allow greater accuracy in early detection of equipment faults.

Recommendation 8: Minimize the impact of maintenance

- System maintenance must be carried out in accordance with a previously scheduled and detailed operation plan.
- Make sure that the periodic technical inspection of power distribution devices are scheduled in accordance with the requirements of the equipment manufacturer or local regulations.
- Maintenance work should be scheduled so as to minimise downtime of HPC infrastructure and critical services.

Recommendation 9: Select the right electricity provider

- In view of the energy market liberalization introduced in the EU, it is possible to negotiate the price of energy with several energy suppliers or consider joining the energy market. Even a small HPC centre with power consumption reaching 1MW is generally classified as a major customer by energy suppliers.
- Reduce the cost of supply charges by ensuring there is no violation of any of the agreed electrical parameters described in an agreement with DNO and energy supplier.
- Limit the usage of harmonics generated by equipment, which back feed the distortions to the DNO network.
- Ensure the distribution network contains elements that help correct reactive power and power factor.

Recommendation 10: Consider regulation and safety

- Regardless of whether the electrical supplies and network elements are operated by the centre staff or are commissioned to an external company, provide comprehensive training for the staff prior to undertaking any work.
- Construction, maintenance and operation should always be executed in accordance with all applicable local regulations.
- Make sure that access to key distribution network is restricted to authorized and properly trained personnel. This may require additional access control devices to restrict access to the distribution network cabinets, rooms etc.

4.2.4 *Redundancy and reliability for an HPC Data Centre*

Abstract

For an HPC centre to operate properly various services, such as energy supply, network connections, data storage systems, building access and automation systems along with computing systems need to be operational. These services in turn provide the HPC centre capacities to deliver services and support for research and computation. Redundancy allows data centres to continue their service in case of unforeseen problems and errors in the infrastructure. Identifying redundancy levels is also an important factor in the design and planning stages of an HPC centre since it will affect the initial investment as well as the data centres physical size and running costs. Future expansion and upgrades might also require additional investment to maintain the same reliability and redundancy levels.

As each HPC centre's location, users, priority services etc. vary, their requirements for redundancy and reliability also vary. The characteristics of the studies and research that rely on computing operated in HPC centres may lead to very different decisions and roadmaps regarding redundancy in the centre.

With the responses to a focused survey received from PRACE consortium partner HPC sites, the effects of redundancy and reliability in terms of positioning of equipment within the system room, energy, cooling and energy efficiency are investigated. Furthermore the paper

includes a cost comparison, effects of unforeseen errors to the data centre and effects of data centre expansions on the reliability and redundancy levels analysed. Best practices are summarized and some recommendations issued.

Recommendations (preliminary version, white paper draft)

Redundancy should be included in the HPC centre design in early stages. This in fact affects the design from the feasibility to the setup of the centre. Each centre should have redundancy policy customized for its location, the equipment used, user profile and centre's aims.

In order to provide continuous operation in the HPC centre reliable equipment and infrastructure are required. Reliable and redundant systems are key to minimal downtime and continuous services delivered to the centre's users.

Prevention and contingency policies for unlikely but possible circumstances such as natural disasters, energy interruptions, security incidents are to be considered.

Optimizing the costs and the levels of redundancy and reliability requires a professional information and experience.

4.2.5 *Security in HPC centres*

This paper is currently being prepared and will be available at [35]. It will analyse the IT and physical security needs of supercomputing centres. The survey related to this white paper can be found in Annex 8.5.

5 HPC Centre Infrastructures - current practices and future trends

5.1 European Workshops on HPC Centre Infrastructures

The series of ***European Workshops on HPC Centre Infrastructures*** is now well established. PRACE-1IP gave the opportunity to accompany the consolidation of these workshops, started during PRACE Preparatory Phase, at the initiative of CSCS (Switzerland), CEA (France) and BAdW/LRZ (Germany).

The Workshops Programme Committee is composed of:

- Ladina Gilly, ETHZ-CSCS (1IP/WP8 Task 2 leader);
- Herbert Huber, BAdW-LRZ (1IP/WP9 former Leader);
- Jean-Philippe Nominé, CEA (1IP/WP8 Leader) ;
- François Robin, CEA (1IP/WP8 Co-leader);
- Dominik Ulmer, ETHZ-CSCS.

The First Workshop took place in Lugano (Switzerland) near CSCS in September 2009, with 50 participants (PRACE Preparatory Phase).

The Second Workshop took place in October 2010, in Dourdan, Paris region (France), near CEA, with 55 participants (prepared during PRACE Preparatory Phase and executed during PRACE-1IP).

The Third Workshop in September 2011, organized by LRZ in Garching (Germany), had 65 participants (prepared and executed during PRACE-1IP).

There is now a strong core of regular Workshops attendees from PRACE and non-PRACE sites but also from the technology side, i.e. providers of both IT and technical equipment. This successfully implements the organizers' willingness to create a stable joint interest group and shows a clear community response to the importance of the issues tackled by the workshops.

The link with PRACE projects still exists through the Programme Committee whose members are also PRACE-1IP WP8 or WP9 members, but there is limited manpower and no specific funding from IP projects for the Workshops organization. This is no real issue since the Programme Committee members' organizations (CEA, CSCS, LRZ) bring in sponsorship and in-kind contributions for the Workshop management and budget balance. This gives good perspectives of independent continuation beyond PRACE IP projects. A fourth Workshop is being considered, possibly taking place early 2013.

5.2 Workshoptopics and audience

Each workshop mixes general presentations of their projects by PRACE or non-European sites with technology presentations by providers and a selection of topical presentations.

The detailed agendas of the three workshops are given in Annex 8.6, 8.7 and 8.8.

5.2.1 *First European Workshop on HPC Centre Infrastructures*

The goal of the first European HPC Infrastructure Workshop was to initiate an exchange of knowledge and experiences. 55 experts came together to discuss topics such as building design, facility management and operation, energy efficiency, cooling technologies and computer cooling designs. Speakers from APC, ASHRAE, Bull, CEA, Cray, CSCS, EYP, Green Grid, IBM, Intel, NCSA, RZ Integral, SGI, SUN, the University of Illinois and the Uptime Institute shaped this event with presentations of high technical quality. Participants also had plenty of opportunities to network and exchange ideas and information.

5.2.2 *Second European Workshop on HPC Centre Infrastructures*

Topics covered were again building design, facility management and operation, energy supply, energy efficiency, cooling technologies and computer cooling designs, with some special focus on novel energy sources and energy market.

Speakers were from EDF, EU Code of Conduct for Data Centres, MTU, GE-Jenbacher, Bull, Cray, HP, IBM, Intel, CEA, LRZ, CSCS, NCSA, ORNL, RIKEN and CSIRO.

A visit of the new CEA facility, TGCC (Très Grand Centre de Calcul du CEA) was organised during the Workshop. Beginning of October 2010 TGCC was just entering the acceptance phase of the building, and about to welcome the first part of the French Tier-0machine "CURIE”.

PRACE representatives from 10 countries and 14 sites as well as representatives from other European sites (CESGA in Spain, Univ. of Gent in Belgium, Univ. of Dresden and DKRZ in Germany, AWE in UK, EPFL in Switzerland) and representatives from world-class supercomputing centres from the USA (ORNL and NCSA), Japan (RIKEN) and Australia (CSIRO) were present.

IT and technical equipment providers represented were MTU, GE-Jenbacher, IBM, Cray, HP, Bull – all of them giving presentations.

With also consultants in the field of data centres (451 Group, EU Code of Conduct), all together 55 attendees took part in the event.

5.2.3 *Third European Workshop on HPC Centre Infrastructures*

The audience of this event mixed again PRACE partners from 9 countries and 11 sites with representatives from other European or American sites or organizations (JRC, CESGA in Spain, Univ. of Ghent in Belgium, TU Dresden, Univ. of Köln, DKRZ and KIT in Germany, EPFL in Switzerland, ECMWF and AWE in UK, Univ. of Vienna in Austria and NCAR in the USA).

Technology and systems providers represented were AMD, Bull, HP, IBM, Intel, Megware and NVIDIA.

Other providers of technical equipment and services represented were ABB/Validus, Emerson, GRC (Green Cooling Revolution), Riedo Networks, deZem, T-Systems, Stulz GmbH and ChristmannInformationstechnik GmbH.

Like each year the programme was composed of different sessions dedicated to different site updates or new projects (PRACE member organizations or others) and technology presentations from IT or technical equipment vendors. This year the special focus was on “building automation”, i.e. all methods and techniques to improve global monitoring and tuning of facilities, especially for energy efficiency.

Speakers were from:

- Sites: LRZ, HLRS, CEA/TGCC, CSC, PSNC, NCAR;
- HPC vendors: AMD, Bull, HP, IBM, Intel, NVIDIA;
- Equipment and services providers : ABB/Validus, Emerson, GRC (Green Cooling Revolution), Riedo Networks, deZem, T-Systems.

A visit of the LRZ facility, whose upgrade is now finished, doubling the five-storey "cube" floor space, was organized during the Workshop.

6 Best practices for the procurement of HPC systems

6.1 General considerations/reminder

Task 8.3 is labelled in the work program as "Best practices for the procurement and installation of HPC systems". Summarizing, the field to be investigated is defined as: "based especially on previous PRACE partners's procurements and procurements carried out during the project duration, including, but not limited to Tier-0". Furthermore: "Risk identification and mitigation will be studied in relation to the different phases of the procurement and commissioning process". The whole task: "carries on and expands the important work started in the PRACE Preparatory Phase project (PP) to come to informed decisions by PRACE as a whole for high-end systems to be acquired". Given the aforementioned main objectives and before we define the methodology to carry on the work started in PRACE-PP, it is useful to summarize the main outcomes of that activity. A first deliverable ([24]), with subject "Procurement strategy", was focused on:

- Issues concerned with the procurement of individual systems;
- A discussion of the strengths and weaknesses of the various EU procurement procedures;
- Capturing lessons learnt within recent procurements by Hosting Partners within PRACE and complementary international procurements.

The procurement of systems was defined in [24](reference to [25]) by its main stages:

- Justification and elaboration – the science and business case;
- Prototype evaluation and tender design to inform the specification of requirements and decision to engage in tender – accepting the market's capacity to supply;
- Investment decision and implementation of the contract;
- Installation;
- Acceptance and pilot use;
- Contract closure;
- Assessment – lessons learnt.

Furthermore the objectives for the procurement of an individual system were defined in [24](reference to [25]) by the following key topics:

- General requirements – positioning the system with respect to a spectrum of application performances, availability and ease-of-use;
- Flexibility of procurement - defining the goal of the procurement from a technical point of view, especially in terms of performances during the lifetime of the system (single delivery vs. phased delivery and related method for evaluating the proposals);
- Costs - defining which parts of the TCO are considered in evaluating the value for money;
- Procurement Process - choosing a national or international procedure and defining its practical implementation (especially regarding schedule);
- Risk minimization - assurance of sustainability, on-going competitiveness of supplier with respect to general HPC market and with respect to the specific regional procurement, warranties and guarantees;

- Maximize value for money - benchmarks and availability acceptance tests. Risk transfer and added value.

Finally, an overview and analysis of a set of recent procurements (2008-2009) was carried out,analysing information based on the following key areas:

- Requirements specification;
- Flexibility of procurement;
- Procurement process (i.e. procurement procedure, with reference to the EU legal framework);
- Benchmarking;
- Evaluation;
- Acceptance;
- Contract.

The framework established by deliverable on “Procurement Strategy” [24] was used as a starting point for this task, and then adapted after having collected information from a number of Tier-0/1 sites about more recent procurements.

On the issue of "risk management", the starting points are the deliverables: "Initial Risk Register" [22] and “Final Risk Register” [23]. The former focused mostly on the risks of supplier default. The latter conducted an analysis of a broader set of risks. It identified, assessed and evaluated risks according to their severity, probability and magnitude, and possible mitigation measures were identified. Forty-three different kinds of risk from the categories "organizational risks" and "technical risks" were assessed in the “Final Risk Register” work [23], the former ones were related to governance, funding and legal structure of the PRACE RI and its supporting institutions. The "technical risks" included considerations at the level of the PRACE RI as a start-up entity. In addition, only risks possibly occurring at the final stage of the procurement process were accounted. Given that the main pillar of risk management is represented by the risk register, we initially concentrated on validating and possibly expanding the register of the risks specific to the procurement process. The aim is to take into account the possible risk sharing between centres and suppliers or technology providers in the context of a procurement that includes an R&D phase. In order to gain insight into procurements with an R&D phase, it may be useful to make an in-depth investigation of such procurements made by PRACE partners. After the involvement of the first partners, we encountered difficulties when asking them for a detailed risk list. Risk management, as we will see, isn't a common topic and attention is focused only on a few key concepts. So we adapted the questionnaire to be more open, allowing each responder to freely expose the field of their own competence.

As a further enhancement to the Preparatory Phase work we introduced pre-commercial procurement, given the increasing importance of the topic for the forthcoming activities related to the PRACE-3IP phase. This is documented in the white paper produced by WP8 on HPC Centre Procurement Best Practice [34].

6.2 Methodology

6.2.1 *Survey*

A procurement process is performed in several stages, with, typically, the following layout:

1. Deciding, organizing and starting the process;
2. Conducting a market information/survey;

3. Preparing the Requests for Proposals (RfP);
4. Preparing benchmarks;
5. Preparing the tender;
6. Conducting the procurement process (including selection of bidder if applicable)
7. Benchmarking;
8. Assessing the proposals and reporting;
9. Decision making;
10. Finalisation of the contract with selected vendors.

Most, if not all, stages are confidential and only little documentation is usually publicly available. Therefore, the only effective way of gathering information is to perform a survey by contacting the right people at the involved sites and asking them a set of questions.

A preliminary questionnaire was prepared for this purpose and was first reviewed by a limited number of sample sites in order to check that the questions were adequate for capturing the main issues. The questions are open but precise enough to make the analysis of the answers useful for the purpose of our work. The questionnaire covered all the main stages of procurement, as well as all the key areas of a comparative analysis using the outcomes of the aforementioned deliverable on "Procurement Strategy"[24].

The preliminary survey was then submitted to several sites involved in recent procurements. The answers to the preliminary questionnaire were analysed and it was then decided to improve the questionnaire before submitting it to more sites refocusing a set of questions and adding five new questions. The full text of the final questionnaire is attached in annex 8.9 (Questionnaire on recent procurements).

The full answers to the survey questionnaires as well as summary tables are reported in a separate confidential document. These include nine full questionnaires from the main tier-0 and tier-1 PRACE partners' sites.

The answers to the final survey were fully analysed in previous confidential deliverable "Consolidated report on Petascale systems and centre assessment" [26], making a logical breakdown into typical HPC system procurement topics:

- Requirements and constraints specification;
- Flexibility and options;
- Costs;
- Internal processes, procedures and risks management;
- Time schedule, contract and negotiation;
- Benchmarking and acceptance;
- Evaluation.

This work led to the production of the publicly available white paper "HPC Systems Procurement Best Practice" available on PRACE web site (www.prace-ri.eu/whitepapers).

6.2.2 *Special considerations regarding risks related to the procurement process*

The risk register developed in aforementioned deliverables "Initial Risk Register" [22] and "Final Risk Register" [23] classifies only few entries explicitly related to the procurement process not related to "technical issues". There are plenty of other risks classified by "Final Risk Register", related to "technical issues" and other not directly mapped to the procurement cycle. After having collected a number of questionnaires we noted that the risk management topic was not sensible to the interviewed and furthermore they were almost uninterested and/or unable to expose some new information about adopted risk management models and/or specific risks to consider. Following that, instead of finding other "risk entries", we considered to compare the work carried out by now with similar experiences or research, in order to gain a more comprehensive point of view. In a recent report (EUR 24229EN, 2010) carried out by an expert group set up by the European Commission Directorate-General for Research, with the subject: "Risk management in the procurement of innovation", various risk management models and practices are analysed and presented. These models range from the simple "implicit risk management", where risks are defined and managed by means of laws and external procedures, to the complex "Integrated risk management model" (see as an example the different phases in Table 13). The big picture derived from the survey conducted in this work package can be analysed with reference to those models.

1. Identify Issues, setting the context.
2. Asses Key Risk Areas
3. Measure likelihood and impact
4. Rank risks
5. Set desired results
6. Develop options
7. Select a strategy
8. Implement the strategy
9. Monitor and evaluate and adjust

Table 13: Steps in Integrated Risk management model

Furthermore the expert panel assessed a classification of risks in (public) procurement, namely: Institutional/societal, Financial, Market, Technological and Other. For a complete definition of each type of risk please refer directly to the original document available on the web. During preliminary phases of the work we analysed the outcomes of "Final Risk Register" and various sources in order to extract a Risk Breakdown Structure (RBS) to be referred to as a starting point for our new classification. The result is depicted in Table 14 and gives the evidence confirmed by the survey: procurer's awareness is well focused on a small subset of the whole field, mainly "technical".

PRACE RI vs. Procurement			
Technical	**External**	**Organizational**	**Management**
Requirements	Subcontractors and suppliers	Dependencies	Estimating
• Contracted technical properties not available; • Data centre doesn't fit requirements; • Power or cooling exceed requirements; • Reliability in requirements; • HW and SW requirements don't match.	• Commercial risks (industrial situation of the vendor) • Technical risks originating at vendor and suppliers • Contractual risks, (contract fulfilment issues related to vendor:	• Coordination with on-going data centre infrastructural projects	• Budget risks (Estimation)

PRACE RI vs. Procurement			
Technical	**External**	**Organizational**	**Management**
	Delivery, Declaration of readiness, Functional properties, Performance commitments, Reliability)		
Technology - Utilization risks (HW and/or SW technology not enough mature)	Regulatory	Resources - Support staff	Planning
Complexity and interfaces - Data centre and infrastructural complexities - Delays due complexity of integration in data centre - Programming styles and software too complex.	Market - Shrinking number of vendors - Staff shortage	Funding - Funding risks (State, National and Supranational)	Controlling
Performances and Reliability - System performance risks (SW and HW related) - Reliability of HW and SW	Environment	Prioritization	Communication
Quality - Final Quality reflect usability	Infrastructures - WAN related risks		

Table 14: RBS with detailed risk

6.3 Conclusion and recommendations

The PRACE project has provided us with a significant evidence base for what works and what does not work in terms of procurements. The key recommendations are captured below, these can be summarized as: know your requirements, know your infrastructure, know the market and plan ahead in terms of investing in software development, and make sure that the user community is ready to exploit the systems. Here the detailed list:

- Requirements: know your requirements and make sure that the market can meet them. We are experiencing an issue in the performance of HPC systems as memory and memory bandwidth are shrinking relative to CPU performance as we move to lower electrical power consumption chips. Quantify required performance as accurately as possible on real workloads.
- Recurrent versus capital: evaluate total lifetime costs in particular power consumption and the trade-off in terms of investment in advanced cooling and electrical supply systems (e.g. combined heat and cooling) versus on-going running costs.
- Infrastructure: plan ahead – power densities within racks are increasing dramatically with the need to move to liquid cooling. This requires specialised mechanical and electrical cooling systems – typically a complete refurbishment of a machine room – with different systems working most optimally at different temperatures requiring different cooling loops.
- Market: timing is becoming more critical – continuously monitor supplier roadmaps to judge when is the best time to go to market and ensure competition.
- Software Development and Testing: invest ahead of the procurement in developing/ adapting the software to run on the new systems. Seek access to prototype systems or software emulators and test and development systems during the operational phase to make sure that you can exploit the system once installed from day one and explore potential performance issues whilst not interrupting normal operations.
- Flexibility: identify cost options as part of the procurement to potentially upgrade the system, bring in new technologies and take advantage of the vendors' longer-term roadmap.

7 Conclusion

This deliverable, complemented by a set of white papers on infrastructure issues and procurement processes, summarizes the results of the work after the 24 months of PRACE-1IP. This effort, based upon work conducted during PRACE Preparatory Phase (PRACE PP WP7), will continue during the PRACE 2nd implementation phase project(PRACE-2IP), except for the study of best practices for the procurement of HPC systems.

This will make possible for PRACE to keep an up to date vision of the HPC market and to address important topics regarding HPC centre infrastructures, focusing on issues regarding operation and thus complementing the work done in the current project that focuses on design, construction or upgrade issues.

Regarding best practices for the procurement of HPC systems, future work will include the joint pre-commercial procurement planned in PRACE 3rd implementation phase (PRACE 3IP). Additional work may be needed in the future in the context of the on-going revision of the public procurement directives that aims at replacing the current directives dating from 2004.

The summary of the work and outcome of the current PRACE-1IP project, coming to an end, are as follows.

Market watch and analysis (task 1)

The aim of the market watch was not only to track the petascale systems and trends based on the Top500 list but also to explore in a systematic way information available on the web (several specific tools were developed for this purpose) and to use the information gathered by PRACE partners. This made possible the collection of details on production systems (installed and running), to identify procured and planned systems that will supposedly enter production from now on and until the next 2-3 years, and to study the major business trends affecting the HPC landscape.

The number of "petascale" systems in production is growing fast; there are more than 20 HPC systems in production that can today be considered as "petascale". The top 10 (last reference November 2011) consists of a mix of hybrid and homogeneous systems as in June 2011. Number 1 is still a non-hybrid system, the K computer in Japan.

More intense changes and evolutions should take place in the next 6 or 12 months, with more systems based on Intel Sandy Bridge or of BlueGene/Q type, or bigger hybrid configurations such as Cray XK6 machines at ORNL or NCSA. The leadership of the USA in the top end systems has clearly been challenged by Japan and China. However, the systems planned in the near future in the USA, including a 20 Petaflop/s system and several 10 Petaflop/s systems, will likely re-establish the dominant position of the USA, or some parity.

Among the aforementioned petascale systems, several ones are Tier-0 PRACE machines – GCS JUGENE and HERMIT in Germany, resp. at FZJ and HLRS, GENCI CURIE in France, at CEA. More PRACE Tier-0, petascale, systems will be fully deployed later in 2012, starting with FERMI at CINECA in Italy, GCS SUPERMUC at LRZ, Germany, consolidating Europe’s position and showing the efforts made by PRACE for maintaining a proper position in the international supercomputer race.

On the longer term, the analysis and extrapolation of current and planned petascale systems is confirming a mix of architectural trends: hybrid/GPU and MPP/clusters, with a steady growth of hybrid systems (GPU and also manycore soon). PRACE now has GPGPU configurations available, but not at large for tier0 capability usage. Real world application readiness for such configurations at large is not obvious to guarantee, nor performance to foresee or guarantee either, but encompassing such architectures in PRACE tier-0 portfolio of systems will be worth considering at some point.

Hardware-software correlation (task 1)

Hardware-software correlation has been considered in cooperation with Work Package 7 (Enabling Petascale Applications), considering programming models, benchmarks and applications (using performance models). The goal was to study the suitability of architectures for real-world applications - which is of key importance in assessing the real value of systems. This work has proved to be very complex and as such it is not definitive, if it ever could be, but it gives first hints on how to proceed.

Among the most interesting result is the identification of programming models and languages available on the current petascale systems. This provides useful input for assessing the portability of programming models and languages on various architectures but doesn't provide at this stage information on the performance issues.

Performance considerations can be addressed by a mix of user feedback from tier-0 actual usage, benchmarking at synthetic and application levels, and possibly performance modelling. No final conclusions were reached on this issue due to the huge complexity of the task. In the future, performance models will likely play an increasing role especially when it comes to anticipating performances on future architectures.

There is no general recipe for application/architecture best mapping, and continuous dialogue between users and computing centre experts is a key factor for finding the best trade-offs in this domain.

Decision making for planning and upgrading HPC centres (task 2)

The work done was entirely focused on producing and disseminating white papers on different topics related to HPC centre infrastructures.

These white papers are meant for quick reference for managers who are not experts in all related technical topics, and can also be considered as check lists for people with more specific technical skills in the concerned fields.

The following topics are addressed in white papers that have been or will soon be published:

- Selecting a Location for an HPC Data Centre
- Cooling – making efficient choices
- Electricity in HPC Centres
- Redundancy and reliability for an HPC Data Centre
- Security in HPC centres

Each white paper is based on the one hand on the expertise of PRACE partners systematically gathered by the means of questionnaires, on the other hand on the presentations given during the HPC centre infrastructure workshops mentioned hereafter.

Updates of these white papers will be considered, in order to keep them up to date and to complement then with further white papers regarding the operation of HPC centre infrastructures during PRACE 2nd implementation phase project.

HPC centre infrastructure - current practices and future trends (task 2)

Three European Workshops on HPC Centre Infrastructures were held resp. in 2009, 2010, and 2011. They were very successful events with 50 to 70 attendees from many PRACE but also other European and non-European sites, from technical equipment providers, and from HPC systems or components vendors.

These events gave participants opportunities to exchange on a diversity of topics. They now have a regular group of attendees, forming an ad hoc 'joint interest group' for HPC infrastructure topics.

More specific issues regarding PRACE partners were discussed during half-day side meetings reserved to PRACE partners. One of the outcomes of these side meetings is the fact that existing or planned HPC centres of PRACE partners are suitable or can evolve in order to host future high end systems.

The series of Workshops is planned to continue next year.

Best practices for the procurement of HPC systems (task 3)

Procuring a high end supercomputer is a complex task because the supercomputer is a complex object by itself; because the procurement process must be selected and then followed carefully, because the selection process is usually based on a mix of large number of requirements and because of the context of fast evolving technology and HPC market landscape. Therefore advice on best practices is of major interest.

While the EU regulation and associated national regulation are, of course, public, very little information is made publicly available on the practical usage of the procedures defined in these regulations by sites conducting such procurements. One of the reasons is procurer and commercial (supplier) confidentiality.

The PRACE context has made possible to share non-commercially confidential information between PRACE partners about procurements of large scale systems. Synthetic information containing a summary of the experience of PRACE partners in this domain has been extractedand made anonymous when relevant for dissemination. This was done through questionnaires that were filled up by PRACE partners for nine large procurements.

Based on this work, a publicly released white paper on procurement best practice, including the issues regarding risk management, was produced. This white paper also contains material on emerging new procedures such as pre-commercial procurement (PCP). PCP is expected to be implemented and experiencedduring PRACE 3rd implementation phase project.

8 Annexes

8.1 « Location » white paper survey

The white paper aims to discuss which important factors need to be considered when selecting a site for a new HPC centre, thus providing managers with a set of key topics that need to be considered when next selecting a site for a new HPC centre. These factors all have an impact on how, and at what cost we will later plan and operate our HPC centre.

In order to collect as much of the knowledge and experience of the PRACE community this survey will be sent to all PRACE sites. It should include the knowledge both of facilities specialists as well as managers.

Please ensure that the survey is, whenever possible, completed by two people: 1 person from the facilities unit of the centre and 1 management member who has or does deal with facilities planning.

The survey is built around a set of open questions to allow you to provide as much input has possible. Please give the details of how things affected you and how you dealt with them.

Questions:

1. Have you been or are you currently involved in the planning, refurbishment or construction of an HPC centre? Please give information of the project you are involved with and in what capacity you contribute to it.

2. Have you been involved in a discussion on whether to extend an existing HPC centre vs. building a new one? If so, what were the arguments for and against each solution and what did you finally elect to do?

3. If the choice fell on a new centre, was there any discussion on building a new building from scratch vs. using an existing structure? If so, what were the arguments for and against each solution and what did you finally elect to do?

4. When selecting a location for a new data centre how free are you would you be in the choice? To what extent do political factors impact the choice vs. other factors? Who makes the final selection decision?

5. What set of criteria did you use to select the site for your new HPC centre? To what extent does the site you chose match these criteria? If you have yet to choose as site, to what extent do you think you will be able to match all criteria?

6. When planning and building your HPC centre, what hurdles or challenges did you come up against, how did they affect the planning or construction project and how did you deal with them? (e.g. zoning limitations, building codes, fire codes, protected natural species, climate, electricity/water/ cooling/land availability, geological quality of terrain, accessibility – for construction and operation, natural hazards/risks, labour regulations, funding, size/dimensions/shape of land, political resistance, resistance from environmental groups or other interest groups, etc.)

7. To what extent did the avoidance of potential hazards impact your choice of location? (e.g. natural hazards, flight paths, pollution, electromagnetic interference, vibration, political climate, etc.)

8. What are the pro and cons of the location you are currently in or are planning to build in? How did they affect the planning or construction project and how did you deal with them? How do they affect operation?

9. Is there one thing you wished you had known/ thought about when selecting the HPC centre location? If so, what is it and how would it have affected your choice?

10. Was there an important lesson you may have learnt/ heard about from a colleague from another HPC centre? If so, what was it and how did it affected them?

Comments: Are there any other topics that you feel are worth mentioning but were not included in the survey above?

8.2 “Cooling” white paper survey

The white paper aims to discuss which important factors need to be considered when selecting a cooling solution for an HPC centre, thus providing managers with a set of key topics that need to be considered when next selecting a site for a new HPC centre. These factors all have an impact on how, and at what cost we plan and operate our HPC centre.
In order to collect as much of the knowledge and experience of the PRACE community this survey will be sent to all PRACE sites. It should include the knowledge both of facilities specialists as well as managers.
Please ensure that the survey is, whenever possible, completed by two people: 1 person from the facilities unit of the centre and 1 management member who has or does deal with facilities planning.
The survey is built around a set of open questions to allow you to provide as much input has possible. Please give the details of how things affected you and how you dealt with them.
If you are involved in planning future data centre or upgrade and currently operating one, please give information for both cases.

Questions:

1. What is your current/planned heat density (e.g. maximum and average Watt/rack or Watt/m2)?

2. What is the maximum capacity of cooling systems in total?

3. Where within (or outside) the data centre are your CRAC/CRAH units located? (i.e. are they located in the computing room or have you foreseen service areas or separate rooms for them?

4. What is your current/planned cooling technology: freon-based/chilled water-based air cooled room (with CRAC units), closed rack cooling (with InROW units), rear-door rack cooling, direct server cooling, ambient outside air, other? Please mention all the technologies you currently use and those you plan to use in future upgrades.

5. Do you have/Are you planning to use a single cooling technology for all devices (only CRAC units, only InROW etc.) or a mixture of solutions? Please give information on what technologies you use and what the rationale behind the choice was.

6. What are the selection criteria for the cooling solution for your data-centre: energy efficiency, capability of cooling X Watt/rack, acquisition price, operational costs, capability of working in your climate, other? For many criteria, please prioritize for each of them.

7. What is the redundancy level of your current/planned cooling solution? What was the reason for choosing the redundancy level?

- active units compressors, chillersetc: e.g. no redundancy, N+1, 2N, other
- infrastructure: e.g. redundant energy distribution panels, redundant water pipes tc.

8. Have you implemented a homogeneous level of cooling equipment redundancy for all devices in the data-centre or are there some parts that are treated in a special way (mission critical servers etc). If there are any, please give the details of which equipment is treated differently and why.

9. Is it possible to scale up the cooling power of your cooling solution or does your cooling solution require some basic infrastructure (e.g. water pipes) to be ready to handle the maximum planned power for the data-centre? To what extent does flexibility and scalability affect your selection of cooling solutions?

10. Do you use any heat re-use techniques: e.g.tri-generation, building heating, commercial heat consumers, other. Do you have any plans for employing such solutions? Is the possibility of heat re-use important enough to influence the location of your future data-centre? Have you implemented any special monitoring systems to control the cooling?

11. Maintenance costs

- What are the annual maintenance costs of your cooling infrastructure?
- What type of support agreement do you have for your cooling infrastructure (e.g.next business day)? Do you have the same service level for all equipment?

Comments: Are there any other topics that you feel are worth mentioning but were not included in the survey above?

8.3 « Electricity» white paper survey

The white paper aims to give recommendations on parameters and issues which have to be taken into account when providing the electrical supply and maintenance of electric distribution network for HPC centre. These factors are playing a crucial role on the cost of operating a HPC centre.
The survey includes questions which will allow collecting knowledge about criteria to choose new location or upgrading an existing one in terms of electricity. It concerns Tier-0 and Tier-1 sites. We appreciate your help by providing us information from facilities engineers and managers. These questions give you a freedom to answer on both options: the case you are involved in upgrading a data centre or planning to build a new one with completely new installation. Please give the details of how things affected you and how you dealt with them.
If you are involved in the process of planning infrastructure of a new data centre or upgrade and currently operating one, please give information for both cases.

Questions:

1. Electric distribution network: short description and simplified diagram (if possible). Please provide following information:
   a. Number of connections (independent energy sources) and voltage levels
   b. Designed power and contracted power for each connection
   c. Issues/restrictions/requirements from DSO
   d. Number, power and voltage level of diesel/gas backup generators
   e. Number and power of transformers (middle voltage to low voltage only)
   f. Measurement devices (energy meters, ammeters, network parameter recorders) and data acquisition
   g. Automatic Transfer Switch and other network automation devices (BMS/BEMS)
   h. Non-standard devices (for example: low-loss transformers with amorphous steel cores, flywheels, etc.)

2. Electric distribution network operation and maintenance
   a. Who operates and provides maintenance for your electric distribution network (own ‘team’, outsourcing, DSO)?
   b. Any kind of issues/limitations/restrictions/annoying things?
   c. What about extension or reconfiguration? Planned/in-progress? Easy? "Don't touch! "?
   d. Do/don’t list based on current experience during operation and maintenance

3. Tariffs and payments
   a. Is your DSO or company related to your DSO also your energy supplier?
   b. Energy price by tariff / negotiated / free market?
   c. Who chooses the tariff plan?
   d. How is the tariff plan chosen?

8.4 « Redundancy and reliability » white paper survey

The white paper aims to discuss which important factors need to be considered when selecting a level of redundancy and reliability for an HPC centre, thus providing managers with a set of key topics that need to be considered when building the next new HPC centre. These factors all have an impact on how, and at what cost we plan the design and acquisition of our HPC centre.

In order for the white paper to contain as much as possible of the knowledge and experience of the PRACE community, this survey will be sent to all PRACE sites. It should include the knowledge both of facilities and certification specialists as well as managers. In this way we will be able to provide managers with a set of key topics that need to be considered when determining the redundancy and reliability requirements for a new HPC centre.

Please ensure that the survey is, whenever possible completed by two people: one person from the facilities unit of the centre and one management member who has dealt or does deal with facilities planning. The survey is built around a set of open questions to allow you to provide as much input as possible. Please give the details of how things affected you and how you dealt with them.

If you are involved in planning a future data centre or upgrade and are currently operating one, please give information for both cases.

Questions:

CRITICAL SERVICES

1. Which services that your centre uses (e.g. cooling, power delivery, networking, computing servers, etc.) are deemed critical?
2. How do you determine which services are critical? (legal, technical reasons, etc.)
3. What are the actions taken to ensure the availability and reliability of these services? Would there be vital factors in decisions regarding these items (e.g. security levels, certification needs, cost of procurement, cost of ownership, ease of maintenance)?

NETWORK

4. Do you have redundant network connections from your HPC centre to the outside world? If so, would you elaborate on the type of connection?

POWER

5. What level of redundancy is built for power in your data centre, in terms of UPS, generators, power grid connections, power lines within the HPC centre?
6. Has the redundancy level for power affected your choices in cooling solutions or layout of the HPC centre?
7. Which devices in cooling systems and power systems are redundant in your HPC centre’s current and/or future setup? Please prioritize the devices according to their redundancy needs. (e.g. CRAC, compressors, coolant lines, power cables etc.)
8. Are there any plans to increase the energy efficiency while maintaining the level of redundancy and reliability?

COOLING

9. Do you have redundancy in cooling for the critical servers or any other section in your data centre? If redundancy in cooling is preferred are there any steps taken to increase the reliability of the critical servers?
10. If any redundancy is implemented on your HPC centre, what is the percentage of cost increase over the conventional non-redundant approach in the areas of cooling, power, networking?
11. If any redundancy is implemented on your HPC centre, how did this affect the layout of the centre?

STORAGE

12. Has your centre implemented or is planning to implement an off-site data storage mirror for disaster recovery?

POLICY

13. Do you use any monitoring systems to track redundancy and reliability of systems used in your current and/or future data centre? If so, please describe what data is collected.
14. Does your current and/or future data centre plan to maintain the same level of redundancy as the data centre expands? How does expandability affect you redundancy plans/strategies?

Comments:

Are there any other topics that you feel are worth mentioning but were not included in the survey above?

8.5 « Security » white paper survey

The white paper aims to discuss which important factors need to be considered when deciding on IT security policies for an HPC Centre.
In order for the white paper to contain as much of the knowledge and experience of the PRACE community this survey will be sent to all PRACE sites. It should include the knowledge both of IT specialists as well as managers. Thus providing managers with a set of key topics that need to be considered regarding IT security issues..
Please ensure that the survey is, whenever possible, completed by two people: 1 person from the IT unit of the centre and 1 management member who has or does deal with facilities planning. The survey is built around a set of open questions to allow you to provide as much input has possible. Please give the details of how things affected you and how you dealt with them in past or current projects.

Questions:

1. Have you got a Security Team / Chief Security Officer?
   a. no
   b. yes – own employees dealing with security issues, but not a separate team/department
   c. yes – own separate security team/department
   How many persons it employs (in FTE)?
   d. yes – outsourcing

2. Has your HPC infrastructure undergone a security test
   a. no
   b. yes, only once/on demand
      - how long ago was the last run?
   c. yes, periodically
      - how often?
      - when was the last run?

3. If any, was the security test performed by:
   a. the people administering the HPC infrastructure
   b. the people from our organization, but not involved with the infrastructure (e.g. a security team)
   c. an external entity

4. What was the scope of the security test? (you can mark more than one option, if applicable)
   a. penetration testing from the Internet (white box or black box?)
   b. penetration testing from the internal network (white box or black box?)
   c. configuration review
   d. other – please describe

5. Have you got an auditing department?
   a. no
   b. yes – own separated department
   c. yes – outsourcing

6. Has your organization undergone a formal (non-technical) security audit?
   a. no
   b. yes, only once/on demand
      - how long ago was the last run?
   c. yes, periodically
      - how often?
      - when was the last run?

7. Are your network interfaces protected by the security systems mentioned below:
   a. network Firewall (please mark if in the HA mode)
   b. network IDS/IPS
   c. DLP software
   d. honeypot
   e. other – please describe

8. Are your particular HPC systems protected by the security systems mentioned below
   a. local IDS/IPS
   b. local firewall
   c. application firewall
   d. antivirus software
   e. other – please describe

9. Have you got an Information Security Policy?
   a. No
   b. yes – the policy exists, but is not based on a known standard (e.g. internal procedures)
   c. yes – the policy exists and is based on a known standard/norm
   - which standard/norm? (e.g. ISO27001)
   - which fields of your organization does it cover?
   - is it certified?

10. How do you manage your systems remotely?
   a. no remote management
   b. remote channel (e.g. ssh/virtual desktop) to a privileged account
   c. remote channel (e.g. ssh/virtual desktop) to a low privileges account and upgrading privileges locally
   d. another way (please describe)

11. Have you got any (and which) countermeasures against a DDoS attack?
   a. no, we don't need them
   b. no, but we are going to implement them
   c. yes, we have procedures how to combat the attack
      Please describe what the procedures are about
   d. yes, we have software/hardware solutions
      Please describe the solutions shortly
   e. yes, we have both organizational and technical countermeasures
      Please describe shortly the procedures and solutions

12. Is your HPC infrastructure separated from the rest of the network?
   a. no separation
   b. logical separation (VLAN) protected by network filters
   c. the HPC part is located in DMZ

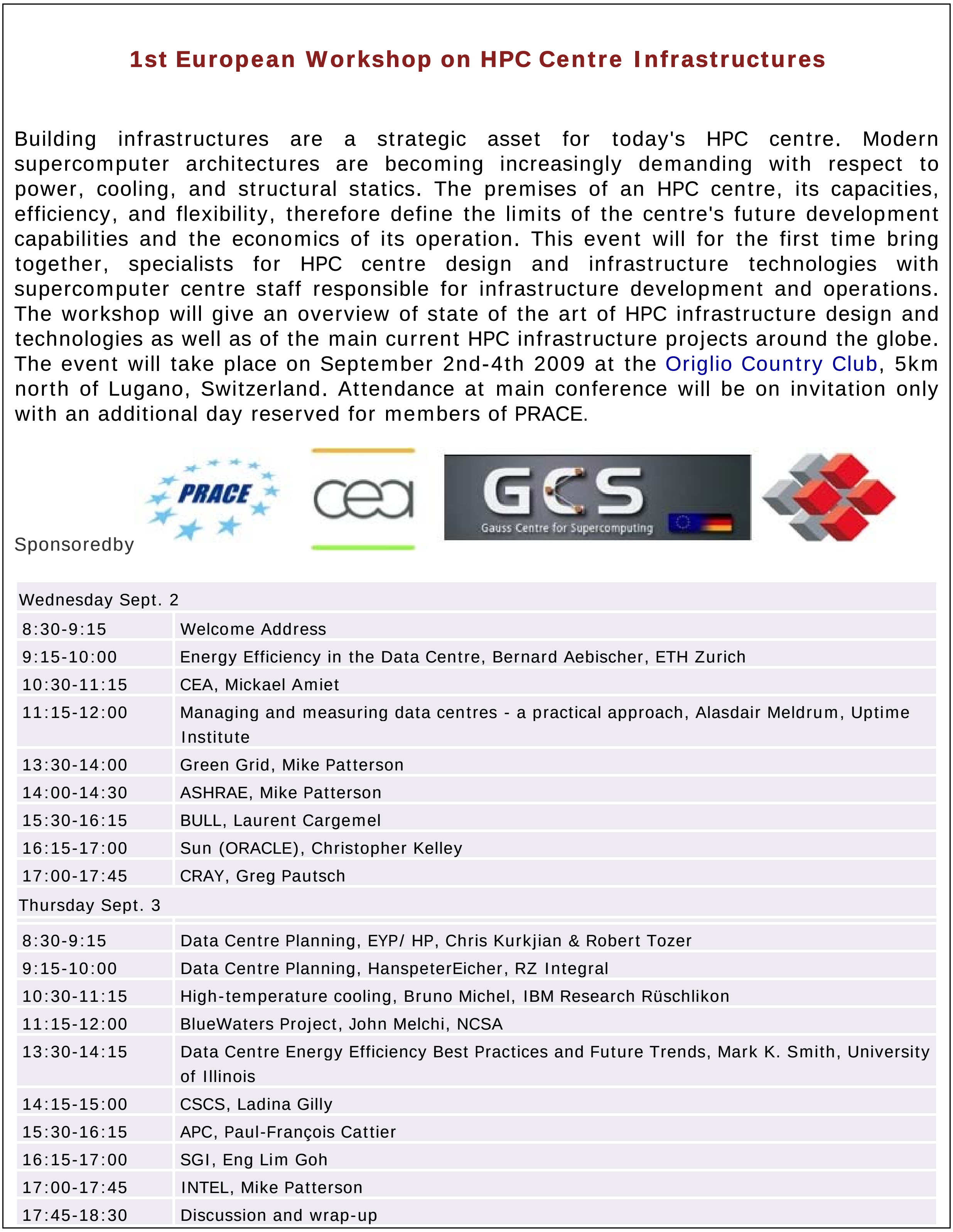
13. Do you use special authentication methods such as one-time-passwords for your users?

14. Do you require a terms of use or equivalent agreement for users to access your center? What are the steps taken if a user has found to breach this agreement?

15. How do you control/restrict physical access to systems (machine rooms, consoles)? (such as tiered security zones, security cameras, biometric identity verification)

Comments:
Are there any other topics concerning IT security that you feel are worth mentioning but were not included in the survey above?

8.6 First European Workshop on HPC Centre Infrastructures -Lugano, Switzerland, 2009



1st European Workshop on HPC Centre Infrastructures

Building infrastructures are a strategic asset for today's HPC centre. Modern supercomputer architectures are becoming increasingly demanding with respect to power, cooling, and structural statics. The premises of an HPC centre, its capacities, efficiency, and flexibility, therefore define the limits of the centre's future development capabilities and the economics of its operation. This event will for the first time bring together, specialists for HPC centre design and infrastructure technologies with supercomputer centre staff responsible for infrastructure development and operations. The workshop will give an overview of state of the art of HPC infrastructure design and technologies as well as of the main current HPC infrastructure projects around the globe. The event will take place on September 2nd-4th 2009 at the Origlio Country Club, 5km north of Lugano, Switzerland. Attendance at main conference will be on invitation only with an additional day reserved for members of PRACE.

Sponsoredby

Wednesday Sept. 2	
8:30-9:15	Welcome Address
9:15-10:00	Energy Efficiency in the Data Centre, Bernard Aebischer, ETH Zurich
10:30-11:15	CEA, Mickael Amiet
11:15-12:00	Managing and measuring data centres - a practical approach, Alasdair Meldrum, Uptime Institute
13:30-14:00	Green Grid, Mike Patterson
14:00-14:30	ASHRAE, Mike Patterson
15:30-16:15	BULL, Laurent Cargemel
16:15-17:00	Sun (ORACLE), Christopher Kelley
17:00-17:45	CRAY, Greg Pautsch
Thursday Sept. 3	
8:30-9:15	Data Centre Planning, EYP/ HP, Chris Kurkjian & Robert Tozer
9:15-10:00	Data Centre Planning, HanspeterEicher, RZ Integral
10:30-11:15	High-temperature cooling, Bruno Michel, IBM Research Rüschlikon
11:15-12:00	BlueWaters Project, John Melchi, NCSA
13:30-14:15	Data Centre Energy Efficiency Best Practices and Future Trends, Mark K. Smith, University of Illinois
14:15-15:00	CSCS, Ladina Gilly
15:30-16:15	APC, Paul-François Cattier
16:15-17:00	SGI, Eng Lim Goh
17:00-17:45	INTEL, Mike Patterson
17:45-18:30	Discussion and wrap-up

8.7 Second European Workshop on HPC Centre Infrastructure -Dourdan, France, 2010

2nd European Workshop on HPC Centre Infrastructures

Dourdan Castle - 13 th centur- Très Grand Centre de Calcul - CEA Bruyères-le-Châtel – 2010

Building infrastructures are a strategic asset for today's HPC centre. Modern supercomputer architectures are becoming increasingly demanding with respect to power, cooling, and structural statics. The premises of an HPC centre, its capacities, efficiency, and flexibility, therefore define the limits of the centre's future development capabilities and the economics of its operation.

This workshop follows the very successful first one held in September 2009 in Lugano, at a time when several important projects of new large HPC facilities arise in Europe, mostly driven by countries and sites which are partners of the PRACE project.

This second event will again bring specialists for HPC centre design and infrastructure technologies together with supercomputer centre staff responsible for infrastructure development and operations. The workshop will give an update on state of the art HPC infrastructure design and technologies as well as the main current HPC infrastructure projects around the globe.

Sponsored by

WednesdayOct 6.

9:00-10:30
- **Welcomeaddress – J.P. Nominé**

Session 1 ***Energy&market issues***
- **Energy & electricity market perspectives – Alain Burtin, EDF**

11:00-12:30 Session 2 ***Energy&Methodology***
- **EC Code of Conduct, Bernard Lecanu**
- **CEA supercomputingcomplexenergy optimisation**
  Mickael Amiet, Jean-Marc Ducos, CEA

13:30-15:00 Session 3 ***Novel energy systems and approaches***
- **Fuel cells – Antonin Guez, MTU**
- **Tri-generation – Joachim Redmer, Bull; Olivier Garcia, GE-Jenbacher**

15:15-16:45 Session 4 ***Facilities I***
- **BAdW/LRZ, Garching – Herbert Huber**
- **ORNL's HPC Infrastructure - Buddy Bland, ONRL**

16:45 -19:30 **CEA TGCC visit (including bus transportation) – CEA/DIF team**

Thursday Oct. 7

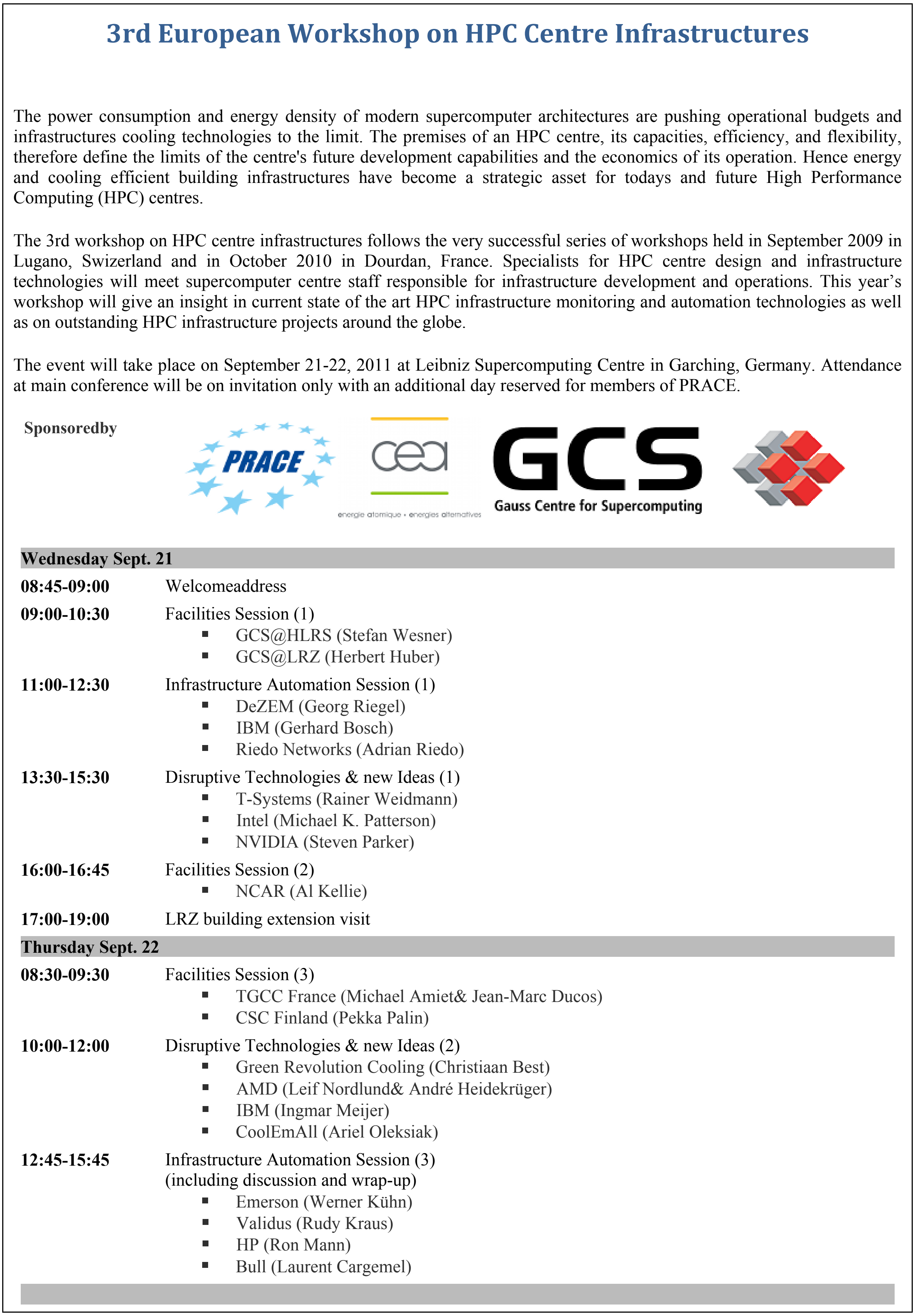
9:00-10:30 Session 5 ***Facilities II***
- **CSCS, Lugano – Ladina Gilly**
- **NCSA – John Melchi**

11:00-12:30 Session 6 ***Facilities III***
- **Facilities of the K computer - Mitsuo Yokokawa, RIKEN**
- **Pawsey Supercomputing Centre and Geothermal integration – Nick Chambers**

13:30-15:00 Session 7 ***IT equipment, HPC/systems***
- **Increased efficiency in power and cooling - Cray – Doug Kelley**
- **The Challenges of today and tomorrow; from Petascale to Exascale**
  Mickael Patterson, Intel

15:30-17:30 Session 8 ***Energyefficiency***
- **Design of Energy Efficient HPC Systems – Gottfried Goldrian, IBM**
- **Infrastructure and energy efficiency in the HPC Data Center-Chris Kurkjian, Pascal Lecoq, HP**
- **Wrapup and conclusion**

8.8 Third European Workshop on HPC Centre Infrastructures -Garching, Germany, 2011



3rd European Workshop on HPC Centre Infrastructures

The power consumption and energy density of modern supercomputer architectures are pushing operational budgets and infrastructures cooling technologies to the limit. The premises of an HPC centre, its capacities, efficiency, and flexibility, therefore define the limits of the centre's future development capabilities and the economics of its operation. Hence energy and cooling efficient building infrastructures have become a strategic asset for todays and future High Performance Computing (HPC) centres.

The 3rd workshop on HPC centre infrastructures follows the very successful series of workshops held in September 2009 in Lugano, Swizerland and in October 2010 in Dourdan, France. Specialists for HPC centre design and infrastructure technologies will meet supercomputer centre staff responsible for infrastructure development and operations. This year's workshop will give an insight in current state of the art HPC infrastructure monitoring and automation technologies as well as on outstanding HPC infrastructure projects around the globe.

The event will take place on September 21-22, 2011 at Leibniz Supercomputing Centre in Garching, Germany. Attendance at main conference will be on invitation only with an additional day reserved for members of PRACE.

Sponsoredby

PRACE

cea energie atomique • energies alternatives

GCS Gauss Centre for Supercomputing

Wednesday Sept. 21

08:45-09:00 Welcomeaddress

09:00-10:30 Facilities Session (1)
- GCS@HLRS (Stefan Wesner)
- GCS@LRZ (Herbert Huber)

11:00-12:30 Infrastructure Automation Session (1)
- DeZEM (Georg Riegel)
- IBM (Gerhard Bosch)
- Riedo Networks (Adrian Riedo)

13:30-15:30 Disruptive Technologies & new Ideas (1)
- T-Systems (Rainer Weidmann)
- Intel (Michael K. Patterson)
- NVIDIA (Steven Parker)

16:00-16:45 Facilities Session (2)
- NCAR (Al Kellie)

17:00-19:00 LRZ building extension visit

Thursday Sept. 22

08:30-09:30 Facilities Session (3)
- TGCC France (Michael Amiet& Jean-Marc Ducos)
- CSC Finland (Pekka Palin)

10:00-12:00 Disruptive Technologies & new Ideas (2)
- Green Revolution Cooling (Christiaan Best)
- AMD (Leif Nordlund& André Heidekrüger)
- IBM (Ingmar Meijer)
- CoolEmAll (Ariel Oleksiak)

12:45-15:45 Infrastructure Automation Session (3)
(including discussion and wrap-up)
- Emerson (Werner Kühn)
- Validus (Rudy Kraus)
- HP (Ron Mann)
- Bull (Laurent Cargemel)

8.9 Questionnaire on recent procurements

Introduction

The aim of this questionnaire is to gather key information about procurements of high-end HPC systems, either recent or planned in the near future. The information will be used to build and improve a set of best practices for future procurements.

Both petascale production systems and prototypes of such systems are targeted by this questionnaire.

Despite the main subject of the survey being the "computational system", the questions also aim to involve the surrounding facilities and the building, if dedicated to the new systems, so feel free to extend your answers to comprise them. Similarly, if the procurement includes peripheral equipment (like storage or visualisation systems), it is desirable to mention it in your answers.

If needed, the collected information can be refined in a next step with a more specific questionnaire/survey or interviews.

Collected information will be treated as confidential and made available only to authorized PRACE project partners as needed.

The questions to be answered are related to the following main categories:

- quite detailed description of the procurement process;
- short description of the risks management during procurement phases;
- descriptionof the HPC system procured or under procurement, either production system or prototype.

In order to set up a "general framework" we consider, from now on, the following main phases for a general procurement process:

1. Deciding, organizing and starting the process;
2. Conducting market information/survey;
3. Preparing Requests for Proposals (RfP);
4. Preparing benchmarks;
5. Preparing tender;
6. Conducting the procurement process (including selection of bidder if applicable)
7. Benchmarking;
8. Assessing the proposals and reporting;
9. Decision making;
10. Finalisation of contract with selected vendors.

Questions

1) How is the decision taken within your organization, to start a procurement process? What are the goals or boundary conditions defined at this stage? Who is involved in this phase?

Consider, amongst others, if there is:

- *a long term roadmap for future systems in your organization;*
- *a formal procedure defined in your Institution for handling procurement, for example defined or inspired by ISO standards.*

2) What is the schedule or time-table for the phases of your procurement process?

Please use, as much as possible, the phases listed here before

3) What is the size and organization of the team on your side? How many people (technical / non-technical / …) are involved in each procurement process phase?

Please use, as much as possible, the phases listed here before.

4) Do you conduct a market survey?

If yes, at what stage of the process? What information is expected from the market survey? Is the market survey useful in your opinion?

5) How do you define and express the requirements of the systems?
Consider especially how you collect and translate into requirements the needs from the users. Consider furthermore as an example the following requirement specification areas: user/application/software, synthetic benchmarks, financial constraints, vendor's future technology road-maps, total or partial power consumption, building/footprint, cooling etc.

6) Which services are considered part of the procurement? How would these services be delivered, what would be the balance between centre and vendor staff?
Consider as example candidate items: installation, maintenance, operations, power supply, training, user support, dismantling, facilities setup, etc.

7) How do you relate system requirements to facility constraints?
The motivation for the question is related to the increasing impact of the upcoming petascale systems in terms of requirements for the surrounding infrastructure.

8) Which costs are considered to drive the procurement process?
Consider as an example if your tender will weigh costs for the offer evaluation. Furthermore, consider if costs are public or not, if your process is cost-driven and if not which component is not cost-driven. Feel free to add whatever you consider relevant.

9) Describe criteria that guided the selection of the type of procedure used for bidder selection.
The answer should consider as an example the following topics: national or international bid, and why; how the contractors are chosen among the available, which criteria; how many vendors are invited during the various phases; which negotiation on final price; etc.
(For the European answerers, please consider types of procedures recalled in footnote[1])

10) Did you find the European and National regulations helped you reach the objectives of the procurement? If not, what were the major obstacles and desirable improvements?

11) How do you structure interaction with vendors during the various phases?
As an example consider which interactions (mail, documents, telephone, meetings …) during survey, selection for RfP, benchmark preparation, etc. You can mention best practices used or identified to avoid situations, which could lead to objections from vendors during the tendering process, if any.

12) To what extent do you allow the vendor degrees of freedom in their offer?

[1] Before to answer to the questions a short explanation about general concepts of European commercial procedure must be clarified. EU Procurement Directives (2004/18/EC) set out the legal framework for public procurements. They apply when public authorities and utilities seek to acquire supplies, services or works and set out procedures which must be followed before awarding a contract when its value exceeds set thresholds, unless it qualifies for a specific exclusion – for example – on the grounds of national security. Our focus is on the following commercial EU procedures, in short:

Open Procedure - Anyone can bid: suppliers respond to a notice in the Official Journal, all interested suppliers will be sent an Invitation to Tender (ITT);

Restricted procedure - The number of tenders (suppliers) may be restricted to at least 5, if available, and only those suitable applicants (assessed by a business questionnaire);

Competitive dialogue procedure - Following an Official Journal EU Contract Notice and a selection process, the authority then enters into dialogue with potential bidders, to develop one or more suitable solutions for its requirements and on which chosen bidders will be invited to tender.

Negotiated procedure - A purchaser may select one or more potential bidders with whom to negotiate the terms of the contract. An example is when, for technical or artistic reasons or because of the protection of exclusive rights, the contract can only be carried out by a particular bidder.

Pre-commercial procurement - was introduced with the intention of driving forward innovation in products and services to address major societal challenges. Typically occupies the product development pipeline starting with solution exploration, moving on to prototyping and then delivering a first limited volume of products/ services (a test series). The value of the pre-commercial procurement contract consists mainly of R&D services and risk-benefit sharing between the procurers and the suppliers.

Consider as an example the followings. Options for offer of reduced/increased system configuration, increased benchmark performance, file system size and performance, memory, interfaces, future technology evaluation systems, additional on-site support. Slippage in delivery for some compensation in system performance.Accommodating technological changes that may be proposed to save money, to improve performance or to save energy or to accommodate increased capability/capacity needs. Feel free to add whatever you consider relevant.

13) To what extent do you require future extensions and upgrades to be built into the offer?
Examples you could consider are the same ones as in previous question Nr 12, but this is the case where the requirements include explicit options to be fulfilled by vendors as part of the offer.

14) How do you evaluate offers? Which evaluation criteria are used?
Please mention if you distinguished mandatory and desirable criteria. How do you score desirable criteria? What kind of weighting do you use for technical criteria? How are the cost(s) taken into account in the evaluation?

15) What is the use of benchmarks in your procurement? How do you perform the benchmarking phase? What kind of benchmarks are used (synthetic, application ..)?
Consider how benchmarks are selected/prepared/executed and which benchmarks are used as selection criteria. Consider if the benchmark set is selected based on user surveys or prior experience, if you collaborate with users to select and prepare the benchmarks or if you prefer to use the outcome of funded projects i.e. PRACE's provided benchmarks. Consider also what is the required scalability of the Benchmarks, duration of execution etc. Consider if you require that the vendor runs the Benchmarks on the installed system.
Consider, if you require performance commitments from vendors on the future system, how do you assess the credibility of the commitments?

16) What is the decision process in your organization for the final selection of the supplier? Is there a body in charge of checking the compliance of the process with European or national regulation?

17) How you manage risks during the whole procurement process? Which risks you would like to be considered or did you consider? Is risk used as a selection criteria? If yes, how?
Try to explain which risks you consider important, how they are identified and how you manage to mitigate them (including in the contracts). Try to enumerate your "top N risk list" ranking them based on "severity of impact (high/medium/low)" and "probability of occurrence (high/medium/low)".
Examples of risks you may consider:

- *risks that may prevent the system from becoming operational: a supplier ceasing to operate or where a system fails to pass its acceptance tests;*
- *risks that may delay the system operation: delays in the production process, delays in sub-contractors roadmaps;*
- *risks that may limit the reliability, availability and serviceability of the system. Lack of key functionality or performance: global parallel file system or power or cooling requirements may exceed expectations or system may not be as reliable as needed;*
- *risks associated with usage/ exploitation of the system: errors in software and hardware or applications performing unexpectedly badly;*

What about the risk the supplier doesn't fulfil its commitments? Do you apply penalties? If yes, what kind and how?

18) What is the outcome of your procurement (what machine did/will you end up with)?
If the system is already procured please indicate the real ones. If the system isn't yet procured, please describe your wishes.

19) Please report here any other information you consider relevant, not covered by previous questions.