SEVENTH FRAMEWORK PROGRAMME
Research Infrastructures

INFRA-2011-2.3.5 – Second Implementation Phase of the European High Performance Computing (HPC) service PRACE

PRACE-2IP

PRACE Second Implementation Project

Grant Agreement Number: RI-283493

D5.1
Preliminary Guidance on Procurements and Infrastructure

Final

Version: 1.0
Author(s): Guillermo Aguirre, BSC
Date: 21.02.2013

Project and Deliverable Information Sheet

PRACE Project	**Project Ref. №: RI-283493**	
	Project Title: PRACE Second Implementation Project	
	Project Web Site: http://www.prace-project.eu	
	Deliverable ID: < **D5.1**>	
	Deliverable Nature: Report	
	Deliverable Level: PU	**Contractual Date of Delivery:** 28/02/2013
		Actual Date of Delivery: 28/02/2013
	EC Project Officer: Leonardo Flores Añover	

Document Control Sheet

Document	**Title: Preliminary Guidance on Procurements and Infrastructure**	
	ID: D5.1	
	Version: <1.0 >	**Status:** ***Final***
	Available at: http://www.prace-project.eu	
	Software Tool: Microsoft Word 2007	
	File(s): D5.1.docx	
Authorship	**Written by:**	Guillermo Aguirre, BSC
	Contributors:	Francois Robin, CEA; Jean-Philippe Nominé, CEA; Ioannis Liabotis, GRNET; Norbert Meyer, PSNC; Radek Januszewski, PSNC; Andreas Johansson, SNIC-LIU; Eric Boyer, CINES; George Karagiannopoulos, GRNET; Marco Sbrighi, CINECA; Vladimir Slavnic, IPB; Gert Svensson, SNIC-KTH
	Reviewed by:	Peter Stefan, NIIF Florian Berberich, PMO & FZJ
	Approved by:	MB/TB

Document Status Sheet

Version	Date	Status	Comments
0.1	16/01/2013	Draft	First outline
0.2	22/01/2013	Draft	Added contributions
0.3	29/01/2013	Draft	Added contributions
0.4	31/01/2013	Draft	Content-complete
0.5	06/02/2013	Draft	Proofread
0.6	08/02/2013	Draft	For internal review
0.7	18/02/2013	Draft	After internal review
1.0	21/02/2013	Final version	

Document Keywords

Keywords:	PRACE, HPC, Research Infrastructure, Petascale

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.eesi-project.eu
[4] http://www.idc.com
[5] http://www.gartner.com
[6] http://www.hpcuserforum.com
[7] https://prace-wiki.fz-juelich.de/bin/view/Prace2IP/WP5/WebHome#Prace2IP_WP5_Web_Utilities (not publicly available)
[8] http://www.netvibes.com/hpc-market-watch#HPC_Market_Watch
[9] http://www.google.com/cse/home?cx=000869963287342404080:2llwvbxdrbo
[10] http://www.top500.org/lists/2012/11/
[11] http://www.extremefactory.com
[12] "High Performance Computing: Europe's place in a Global Race" - Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the regions, 15.2.2012, COM(2012)15
[13] "Europe achieving leadership in HPC" http://www.etp4hpc.eu/news-and-events/docs/HPC%20leadership%20final%20121101.pdf
[14] http://mpc.sourceforge.net/
[15] http://eehpcwg.lbl.gov

List of Acronyms and Abbreviations

AAA	Authorization, Authentication, Accounting.
ACF	Advanced Computing Facility
ADP	Average Dissipated Power
AISBL	Association Internationale Sans But Lucratif (legal form of the PRACE-RI)
AMD	Advanced Micro Devices
APGAS	Asynchronous PGAS (language)
API	Application Programming Interface
APML	Advanced Platform Management Link (AMD)
ASIC	Application-Specific Integrated Circuit
ATI	Array Technologies Incorporated (AMD)
BAdW	Bayerischen Akademie der Wissenschaften (Germany)
BCO	Benchmark Code Owner
BSC	Barcelona Supercomputing Center (Spain)
CAF	Co-Array Fortran
CAL	Compute Abstraction Layer
CCE	Cray Compiler Environment
ccNUMA	cache coherent NUMA
CEA	Commissariat à l'Energie Atomique et aux Energies Alternatives (represented in PRACE by GENCI, France)
CINECA	Consorzio Interuniversitario, the largest Italian computing centre (Italy)
CINES	Centre Informatique National de l'Enseignement Supérieur (represented in PRACE by GENCI, France)
CLE	Cray Linux Environment
CPU	Central Processing Unit
CSC	Finnish IT Centre for Science (Finland)
CSCS	The Swiss National Supercomputing Centre (represented in PRACE by ETHZ, Switzerland)
CUDA	Compute Unified Device Architecture (NVIDIA)
DARPA	Defense Advanced Research Projects Agency
DDR	Double Data Rate
DEISA	Distributed European Infrastructure for Supercomputing Applications. EU project by leading national HPC centres.
DIMM	Dual Inline Memory Module
DMA	Direct Memory Access
DP	Double Precision, usually 64-bit floating point numbers
DRAM	Dynamic Random Access memory
EC	European Community
EESI	European Exascale Software Initiative
EFlop/s	Exa (= 10^{18}) Floating point operations (usually in 64-bit, i.e. DP) per second, also EF/s
EPCC	Edinburg Parallel Computing Centre (represented in PRACE by EPSRC, United Kingdom)
EPSRC	The Engineering and Physical Sciences Research Council (United Kingdom)
ESFRI	European Strategy Forum on Research Infrastructures; created roadmap for pan-European Research Infrastructure.
ETHZ	Eidgenössische Technische Hochschule Zuerich, ETH Zurich (Switzerland)
ETP	European Technology Platform

FHPCA	FPGA HPC Alliance
FP	Floating-Point
FPGA	Field Programmable Gate Array
FPU	Floating-Point Unit
FZJ	Forschungszentrum Jülich (Germany)
GASNet	Global Address Space Networking
GB	Giga (= 2^{30} ~ 10^9) Bytes (= 8 bits), also GByte
Gb/s	Giga (= 10^9) bits per second, also Gbit/s
GB/s	Giga (= 10^9) Bytes (= 8 bits) per second, also GByte/s
GCS	Gauss Centre for Supercomputing (Germany)
GDDR	Graphic Double Data Rate memory
GÉANT	Collaboration between National Research and Education Networks to build a multi-gigabit pan-European network, managed by DANTE. GÉANT2 is the follow-up as of 2004.
GENCI	Grand Equipement National de Calcul Intensif (France)
GFlop/s	Giga (= 10^9) Floating point operations (usually in 64-bit, i.e. DP) per second, also GF/s
GHz	Giga (= 10^9) Hertz, frequency =10^9 periods or clock cycles per second
GigE	Gigabit Ethernet, also GbE
GLSL	OpenGL Shading Language
GNU	GNU's not Unix, a free OS
GPGPU	General Purpose GPU
GPU	Graphic Processing Unit
GWU	George Washington University, Washington, D.C. (USA)
HDD	Hard Disk Drive
HE	High Efficiency
HET	High Performance Computing in Europe Taskforce. Taskforce by representatives from European HPC community to shape the European HPC Research Infrastructure. Produced the scientific case and valuable groundwork for the PRACE project.
HMPP	Hybrid Multi-core Parallel Programming (CAPS enterprise)
HP	Hewlett-Packard
HPC	High Performance Computing; Computing at a high performance level at any given time; often used synonym with Supercomputing
HPCC	HPC Challenge benchmark, http://icl.cs.utk.edu/hpcc/
HPCS	High Productivity Computing System (a DARPA program)
HPL	High Performance LINPACK
HT	HyperTransport channel (AMD)
HWA	HardWare accelerator
IB	InfiniBand
IBA	IB Architecture
IBM	Formerly known as International Business Machines
ICE	(SGI)
IDRIS	Institut du Développement et des Ressources en Informatique Scientifique (represented in PRACE by GENCI, France)
IEEE	Institute of Electrical and Electronic Engineers
IESP	International Exascale Project
I/O	Input/Output
ISC	International Supercomputing Conference; European equivalent to the US based SC0x conference. Held annually in Germany.
JSC	Jülich Supercomputing Centre (FZJ, Germany)

KB	Kilo (= 2^{10} ~10^3) Bytes (= 8 bits), also KByte
KTH	Kungliga Tekniska Högskolan (represented in PRACE by SNIC, Sweden)
LINPACK	Software library for Linear Algebra
LLNL	Laurence Livermore National Laboratory, Livermore, California (USA)
LRZ	Leibniz Supercomputing Centre (Garching, Germany)
MB	Mega (= 2^{20} ~ 10^6) Bytes (= 8 bits), also MByte
MB/s	Mega (= 10^6) Bytes (= 8 bits) per second, also MByte/s
MFlop/s	Mega (= 10^6) Floating point operations (usually in 64-bit, i.e. DP) per second, also MF/s
MHz	Mega (= 10^6) Hertz, frequency =10^6 periods or clock cycles per second
MIPS	Originally Microprocessor without Interlocked Pipeline Stages; a RISC processor architecture developed by MIPS Technology
Mop/s	Mega (= 10^6) operations per second (usually integer or logic operations)
MPI	Message Passing Interface
MPP	Massively Parallel Processing (or Processor)
MPT	Message Passing Toolkit
NCF	Netherlands Computing Facilities (Netherlands)
NDA	Non-Disclosure Agreement. Typically signed between vendors and customers working together on products prior to their general availability or announcement.
NFS	Network File System
NIC	Network Interface Controller
NUMA	Non-Uniform Memory Access or Architecture
OpenCL	Open Computing Language
OpenGL	Open Graphic Library
Open MP	Open Multi-Processing
OS	Operating System
PCIe	Peripheral Component Interconnect express, also PCI-Express
PCI-X	Peripheral Component Interconnect eXtended
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
PGI	Portland Group, Inc.
pNFS	Parallel Network File System
PRACE	Partnership for Advanced Computing in Europe; Project Acronym
PSNC	Poznan Supercomputing and Networking Centre (Poland)
QDR	Quad Data Rate
RAM	Random Access Memory
RISC	Reduce Instruction Set Computer
RPM	Revolution per Minute
SARA	Stichting Academisch Rekencentrum Amsterdam (Netherlands)
SAS	Serial Attached SCSI
SATA	Serial Advanced Technology Attachment (bus)
SDK	Software Development Kit
SGI	Silicon Graphics, Inc.
SHMEM	Share Memory access library (Cray)
SIMD	Single Instruction Multiple Data
SM	Streaming Multiprocessor, also Subnet Manager
SMP	Symmetric MultiProcessing
SNIC	Swedish National Infrastructure for Computing (Sweden)

SP	Single Precision, usually 32-bit floating point numbers
SRA	Strategic Research Agenda
SSD	Solid State Disk or Drive
STFC	Science and Technology Facilities Council (represented in PRACE by EPSRC, United Kingdom)
STRATOS	PRACE advisory group for STRAtegic TechnOlogieS
TB	Tera (= 240 ~ 1012) Bytes (= 8 bits), also TByte
TCO	Total Cost of Ownership. Includes the costs (personnel, power, cooling, maintenance, ...) in addition to the purchase cost of a system.
TDP	Thermal Design Power
TFlop/s	Tera (= 1012) Floating-point operations (usually in 64-bit, i.e. DP) per second, also TF/s
Tier-0	Denotes the apex of a conceptual pyramid of HPC systems. In this context the Supercomputing Research Infrastructure would host the Tier-0 systems; national or topical HPC centres would constitute Tier-1
UFM	Unified Fabric Manager (Voltaire)
UPC	Unified Parallel C
UV	Ultra Violet (SGI)
VHDL	VHSIC (Very-High Speed Integrated Circuit) Hardware Description Language

Executive Summary

The PRACE-2IP Work Package 5 (WP5), "Best Practices for HPC Systems Commissioning", has two objectives:

- Procurement, independent vendor relations and market watch (task 1)
- Best practices for HPC Centre Infrastructures (task 2)

This Work Package builds on and expands the important work started in the PRACE Preparatory Phase project (PRACE-PP WP7) and continued through PRACE 1st Implementation Phase (PRACE-1IP WP8), which have all sought to reach informed decisions within PRACE as a whole on the acquisition and hosting of HPC systems and infrastructure.

WP5 provides input for defining and updating procurement plans and strategies through the sharing of the state of the art and best practices in procuring and operating production HPC systems.

Task 1 – **Assessment of petascale systems** – has performed a continuous market watch and analysis of trends in petascale HPC systems worldwide. The information, collected from public sources and industry conferences, is presented through comparisons and graphs that allow an easier and quicker examination of trends for different aspects of top-level HPC systems. Specific areas of interest are analysed in depth in terms of the market they belong to and the general HPC landscape, with a particular emphasis on the European point of view.

As well as the general market analysis, this task also describes the link between hardware and software (in collaboration with software-specific work packages WP11 and WP12), providing information on the supported interfaces (programming models), benchmark results and user requirements.

Task 2 – **Best practices for designing and operating power efficient HPC centre infrastructures** – has continued the production of white papers which explore specific topics related to HPC data centre design and operation, with input from PRACE members. It has also analysed the current state of the art in cooling and power efficient operating of HPC infrastructure.

In addition, this task also includes the organization of the fourth edition of the "European Workshop on HPC Centre Infrastructures", which will be hosted by CSCS in Lugano in the spring of 2013.

1 Introduction

D5.1 – Preliminary Guidance on Procurements and Infrastructure – is the first of two deliverables to be produced by WP5, which started its activity in M12 to not overlap with similar work done in PRACE-1IP WP8. Its aim is to serve as the foundation for the final report due in M24.

This document is organised in 5 main chapters, in addition to this introduction (Chapter 1) and to the conclusions (Chapter 7):

- Chapter 2 - an update on the petascale HPC market and its evolution, as well as the sources and tools used to collect this information and more detailed insight of specific players: HPC vendors and accelerator manufacturers. It also presents the European perspective on HPC and the position of PRACE in the global context.
- Chapter 3 – an analysis of the relationship between hardware and its usage within PRACE Tier-0 systems, from the programming models used to the specific applications and their execution.
- Chapter 4 – a report on current trends in HPC energy efficiency as presented at the latest Supercomputing Conference (SC12).
- Chapter 5 – a summary of the conclusions and recommendations presented in the white paper produced on the topic of HPC Site Security.
- Chapter 6 – an outline of the work done to organise the 4th European Workshop on HPC Centre Infrastructure.

Some additional details are provided in annex:

- Annex 1 – The questionnaire sent to PRACE member sites for the “HPC Centre Security” whitepaper.

2 Assessment of petascale systems

Since the advent of petascale computing in 2008 the size and complexity of this high-end HPC market has continually increased. In the past four years we have seen how Roadrunner, the first computer in the world to reach a peak performance of 1 PFlop/s, has been joined by another 31 systems from around the globe in the exclusive "petascale club". These 32 supercomputers make use of diverse strategies to leverage state of the art technology for achieving their peak performance while optimizing resources: from the use of accelerators and lightweight many-core processors to reduce power consumption, to the use of novel interconnects with improved bandwidth and latency for reducing bottlenecks.

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 three sections:

- **Market Watch and Analysis** outlines the current situation in petascale HPC by providing a detailed look at both the present-day petascale systems and their evolution in time.
- **Business Analysis** describes the general trends observed in the HPC market, as well as a more in-depth look at some of its most important submarkets: vendors and accelerator manufacturers.
- **European HPC Ecosystem in a Global Context** provides a glimpse of the European position with respect to the world, and describes initiatives taken to maintain Europe at the forefront of HPC.

2.1 Market Watch and Analysis

This section contains a comprehensive analysis of the high-end HPC market, specifically limited to systems with a peak performance of at least 1 PFlop/s. This examination combines both an exhaustive description of the current 32 publicly recognized petascale systems in the world as well as an overview of their evolution in time, and includes:

- A catalogue of publicly available **sources** from which the raw data for the analysis has been extracted, as well as tools developed specifically for this purpose.
- A **snapshot** of current petascale systems as presented in the November 2012 edition of the Top500 List.
- Statistics and graphs used to compare the characteristics of the supercomputers contained in the snapshot, providing a **static analysis** of the present market.
- Time-based statistics and graphs provide insight into the evolution and trends, in the form of a **dynamic analysis** of the petascale market.

2.1.1 *Sources*

All the raw data used to produce the analyses found in this chapter have been collected from a variety of public sources available on the Internet, and reorganized in a structured manner in the PRACE internal wiki for use by PRACE and its members. This section provides links and

descriptions of the main sources of information used for this purpose, as well as tools that have been specifically developed to aid in this data-collection process.

The sources presented below are similar to the ones that were given in PRACE-1IP WP8 deliverables, correcting any out-dated links in cases that those existed and adding new links based on new findings (i.e. new Top500 lists). Finally at the end of the section we document the several developments that help the identification of information relevant to the market watch, namely the internal Market Watch wiki page that contains links to the Netvibes [1] feeds aggregator and the Market Watch Google Custom Search Engine (CSE) [2].

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

1. **HPC related electronic publications / web sites**: Those 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 users demand. Those 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**: Those 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 web sites of the vendors, contain a variety of information on the new technologies developed and deployed by them. They are presented as product documentation, white papers, press releases, etc. Additionally, 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 issue. The vendor specific web sites offer mostly the vendor's point of view on the HPC market.
4. **Funding agencies web sites**: Those web sites are maintained by various funding agencies around the world. This is where someone can find information on new or planned procurements via press releases or RFIs/RFPs 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 three categorizations that were described above.

HPC related electronic publications and web sites

- http://www.top500.org/ - The Top 500 supercomputer sites, publishes the top 500 list of general purpose systems that are in common use for high-end applications. The present Top 500 list, lists computers ranked by their performance on the LINPACK Benchmark. The list is updated half-yearly and, in this way there is track of the evolution of computers.
- http://www.green500.org/ - 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 uses "MFlop/s-per-Watt" as its ranking metric (based on LINPACK execution), while other lists are also published based on community feedbacks.

- http://www.hpcwire.com/ - HPCWire is an on line publication devoted to HPC news. It is one of the most popular on line publications for people involved in High Performance Computing. The news, are categorized in 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.

A few other electronic publications that can be used for searching for information on current and future HPC systems are:

- HPC Inside – http://insidehpc.com/
- Scientific Computing.COM - http://www.scientific-computing.com/
- Microprocessor report - http://www.mdronline.com/
- Supercomputing online - http://www.supercomputingonline.com/

In this category we can also add the European Exascale Software Initiative [3]. 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 - from petascale 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.

Firms such as IDC [4] or GARTNER [5] are also 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 selected and synthetic information (e.g. IDC and HPC User Forum [6], or regular market updates with some predictions and forecast).

Computing centre websites

The list of computing centre web sites, obtained from the November 2012 Top 500 list, is presented in Table 1.

System	Site	Web Address
Titan	DOE/SC/Oak Ridge National Laboratory United States	http://www.olcf.ornl.gov/titan/
Sequoia	DOE/NNSA/LLNL United States	https://asc.llnl.gov/computing_resources/sequoia/index.html
K computer	RIKEN Advanced Institute for Computational Science (AICS) Japan	http://www.aics.riken.jp/en/kcomputer/
Mira	DOE/SC/Argonne National Laboratory United States	https://www.alcf.anl.gov/mira
JUQUEEN	Forschungszentrum Juelich (FZJ) Germany	http://www.fz-juelich.de/ias/jsc/EN/Expertise/Supercomputers/JUQUEEN/JUQUEEN_node.html
SuperMUC	Leibniz Rechenzentrum (LRZ) Germany	http://www.lrz.de/services/compute/supermuc/
Stampede	Texas Advanced Computing Center - University of Texas United States	http://www.tacc.utexas.edu/stampede
Tianhe-1A	National Supercomputing Center in Tianjin China	http://www.nscc-tj.gov.cn/en/

System	Site	Web Address
Fermi	CINECA Italy	http://www.hpc.cineca.it/content/fermi-reference-guide
Curie TN	CEA/TGCC-GENCI France	http://www-hpc.cea.fr/en/complexe/tgcc-curie.htm
Nebulae	National Supercomputing Centre in Shenzhen (NSCS) China	http://www.nsccsz.gov.cn
Yellowstone	NCAR (National Center for Atmospheric Research) United States	https://www2.cisl.ucar.edu/resources/yellowstone/hardware
Pleiades	NASA/Ames Research Center/NAS United States	http://maeresearch.ucsd.edu/beg/pleiades.php
TSUBAME 2.0	GSIC Center, Tokyo Institute of Technology Japan	http://tsubame.gsic.titech.ac.jp/en
Cielo	DOE/NNSA/LANL/SNL United States	http://www.lanl.gov/orgs/hpc/cielo/index.shtml
Hopper	DOE/SC/LBNL/NERSC United States	http://www.nersc.gov/users/computational-systems/hopper/
Tera-100	Commissariat a l'Energie Atomique et aux Energies Alternatives (CEA) France	http://www-hpc.cea.fr/fr/complexe/docs/T100.htm
Oakleaf-FX	Information Technology Center, University of Tokyo Japan	http://www.cc.u-tokyo.ac.jp/system/fx10/
Roadrunner	DOE/NNSA/LANL United States	http://www.lanl.gov/roadrunner/
DiRAC	University of Edinburgh United Kingdom	http://www.stfc.ac.uk/Our+Research/24711.aspx

Table 1: HPC computing centre URLs

Vendor 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 (SC12) in Salt Lake City, 199 industry exhibitors showed their products.

The following list of vendors is based on the vendors that supplied the most powerful 50 systems of the November 2012 Top 500 list. Note that National University of Defense 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.

- Appro International - http://www.appro.com/
- Bull SA - http://www.bull.com/extreme-computing/index.html
- Cray Inc. - http://www.cray.com/Products/Products.aspx
- Dawning - http://www.sugon.com/chpage/c1/
- Dell - http://content.dell.com/us/en/enterprise/hpcc.aspx?cs=555
- Fujitsu - http://www.fujitsu.com/global/services/solutions/tc/hpc/products/
- Hewlett-Packard - http://h20311.www2.hp.com/hpc/us/en/hpc-index.html
- IBM - http://www-03.ibm.com/systems/deepcomputing/
- NEC - http://www.necam.com/hpc/
- NVIDIA - http://www.nvidia.com/object/tesla_computing_solutions.html
- SGI - http://www.sgi.com/
- Oracle - http://www.oracle.com/us/products/servers-storage/index.html

- Raytheon - http://www.raytheon.com/capabilities/products/hpc/
- T-Platforms - http://www.t-platforms.ru/new/
- TYAN - http://www.tyan.com/products.aspx

Funding agencies web sites

Table 2 presents the web addresses 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://science.energy.gov/
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

Table 2: Funding agencies' URLs

Market Watch Tools

A set of tools have been deployed within PRACE-1IP WP8, and maintained in PRACE-2IP WP5, in order to take advantage of the above collection of links for the Market watch. Those tools would facilitate 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. Both the tools as well as an up to date list of the web sources that appear in the previous sections are available to PRACE members in the internal wiki [7].

- **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. An aggregator reduces the effort and therefore the time needed to check a big list of web sites for updates creating a single page where information from pre-selected web content can be viewed. For the purposes of the Market Watch activity of WP5 the Netvibes [1] aggregator has been maintained and used, and can be accessed freely. This page [8] allows us to easily monitor the HPC market news periodically without the need to visit all the web sites that have been collected as our sources. Currently the Netvibes page contains 16 news or twitter feeds from the list of our sources.
- **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. That way CSE provides only relevant search results, thus speeding the process of searching information that is needed. Within WP5 we have created a Market Watch CSE that contains all sites that are relevant to the activity, which can be accessed directly from a Google URL [9].

2.1.2 *Snapshot*

According to the latest version (November 2012) of the Top500 list of worldwide supercomputers [10], presented at SC12 on November 12th 2012, there are currently 32 systems with a peak performance of 1 PFlop/s or more. These systems, described briefly in Table 3, will be used in the subsequent comparison and analysis of the following section.

This relatively small subset provides a glimpse of the requirements and techniques used to reach petascale performance, as well as the market situation and trends. By comparing the architectural characteristics of these machines, we can classify them into three broad categories:

- Accelerated: use co-processors to handle part of the load (in red)
- Lightweight: use many low-power RISC processors (in green)
- Traditional: use only standard high-performance processors (in blue)

System	Site (Country)	Model (processor / accelerator)	LINPACK / peak (PFlop/s)
Titan	ORNL (USA)	Cray XK7 (AMD Opteron / NVIDIA Tesla)	17.59 / **27.11**
Sequoia	LLNL (USA)	IBM Blue Gene/Q (IBM PowerPC)	16.33 / **20.13**
K Computer	RIKEN (Japan)	Fujitsu Cluster (Fujitsu SPARC64)	10.51 / **11.28**
Mira	ANL (USA)	IBM Blue Gene/Q (IBM PowerPC)	8.16 / **10.07**
JUQUEEN	FZJ (Germany)	IBM Blue Gene/Q (IBM PowerPC)	4.14 / **5.03**
SuperMUC	LRZ (Germany)	IBM iDataPlex (Intel Xeon)	2.90 / **3.19**
Stampede	TACC (USA)	Dell PowerEdge (Intel Xeon / Intel Xeon Phi)	2.66 / **3.96**
Tianhe-1A	NSCT (China)	NUDT YH MPP (Intel Xeon / NVIDIA Tesla)	2.57 / **4.70**
Fermi	CINECA (Italy)	IBM Blue Gene/Q (IBM PowerPC)	1.73 / **2.10**
DARPA TS	IBM DE (USA)	IBM Power 775 (IBM POWER7)	1.52 / **1.94**
Curie thin nodes	TGCC (France)	Bull B510 (Intel Xeon)	1.36 / **1.67**
Nebulae	NSCS (China)	Dawning TC3600 (Intel Xeon / NVIDIA Tesla)	1.27 / **2.98**
Yellowstone	NCAR (USA)	IBM iDataPlex (Intel Xeon)	1.26 / **1.50**
Pleiades	NAS (USA)	SGI Altix ICE (Intel Xeon)	1.24 / **1.73**
Helios	IFERC (Japan)	Bull B510 (Intel Xeon)	1.24 / **1.52**
Blue Joule	STFC (UK)	IBM Blue Gene/Q (IBM PowerPC)	1.21 / **1.47**
Tsubame 2.0	GSIC (Japan)	NEC/HP ProLiant (Intel Xeon / NVIDIA Tesla)	1.19 / **2.29**
Cielo	LANL (USA)	Cray XE6 (AMD Opteron)	1.11 / **1.37**
Hopper	NERSC (USA)	Cray XE6 (AMD Opteron)	1.05 / **1.29**
Tera-100	CEA (USA)	Bull S6010/S6030 (Intel Xeon)	1.05 / **1.26**
Oakleaf-FX	SCD (Japan)	Fujitsu PRIMEHPC (Fujitsu SPARC64)	1.04 / **1.14**
Roadrunner	LANL (USA)	IBM TriBlade (AMD Opteron / IBM PowerXCell)	1.04 / **1.38**
DiRAC	EPCC (UK)	IBM Blue Gene/Q (IBM PowerPC)	1.04 / **1.26**
Anonymous	NCI (Australia)	Fujitsu PRIMERGY (Intel Xeon)	0.98 / **1.11**
Kraken XT5	NICS (USA)	Cray XT5 (AMD Opteron)	0.92 / **1.17**
Lomonosov	RCC (Russia)	T-Platforms T-Blade (Intel Xeon / NVIDIA Tesla)	0.90 / **1.70**
Hermit	HLRS (Germany)	Cray XE6 (AMD Opteron)	0.83 / **1.04**
Sunway BlueLight	NSC (China)	Sunway Cluster (ShenWei SW1600)	0.80 / **1.07**
Tianhe-1A HS	NSCCH (China)	NUDT YH MPP (Intel Xeon / NVIDIA Tesla)	0.77 / **1.34**

System	Site (Country)	Model (processor / accelerator)	LINPACK / peak (PFlop/s)
Mole-8.5	IPE (China)	Tyan FT72-B7015 (Intel Xeon / NVIDIA Tesla)	0.50 / **1.01**
Anonymous	Geoscience (USA)	HP ProLiant (Intel Xeon / NVIDIA Tesla)	0.46 / **1.14**
SANAM	KACST (Saudi Arabia)	Adtech custom (Intel Xeon / AMD FirePro)	0.42 / **1.10**

Table 3: Snapshot of current petascale systems

2.1.3 *Static Analysis*

Statistical analysis of the features of the set of petascale systems indicated in the Market Watch can offer relevant conclusions regarding the current situation. By analysing each characteristic independently and then merging the resulting conclusions, the market can be described in much more detail, providing a better understanding of the underlying environment.

Year of construction

Peak performance of 1 PFlop/s was reached for the first time in 2008 (in general purpose publicly listed computers), yet none of the current systems date that far back (only Jaguar and Roadrunner were petascale then, and they have both been updated since). In fact, almost half of the systems in the market watch (around 47%) were built or updated in 2012, while only a meagre 3% still remain from 2009.

With regards to the architecture, it is interesting to see that lightweight systems are exclusively in the 2011-2012 range (the first lightweight petascale system was JUGENE in 2009, but it has now been replaced by JUQUEEN).

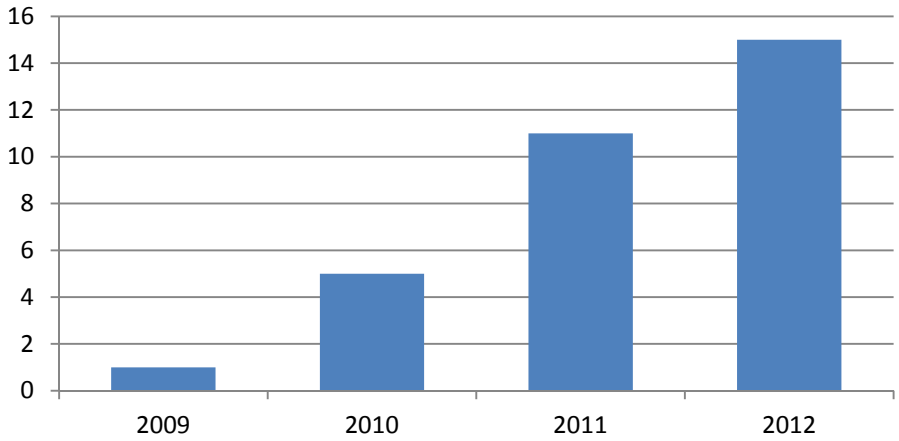

Figure 1: Petascale systems by year of construction

Country

Almost two fifths of all the petascale systems are installed in the USA (37.5%), including the top two performers (Titan and Sequoia). China, whose fastest system is now in eighth place in the November 2012 Top500 list, is the first runner-up in terms of market share, although at a considerable distance from the USA with only about 16%. Japan is the next big player, with 4 petascale systems giving them 12.5% of the market, and Germany follows at 9.4% share. France and UK each have 2 petascale systems (6.25% market share), and Italy, Russia, Australia, and Saudi Arabia complete the list with one each.

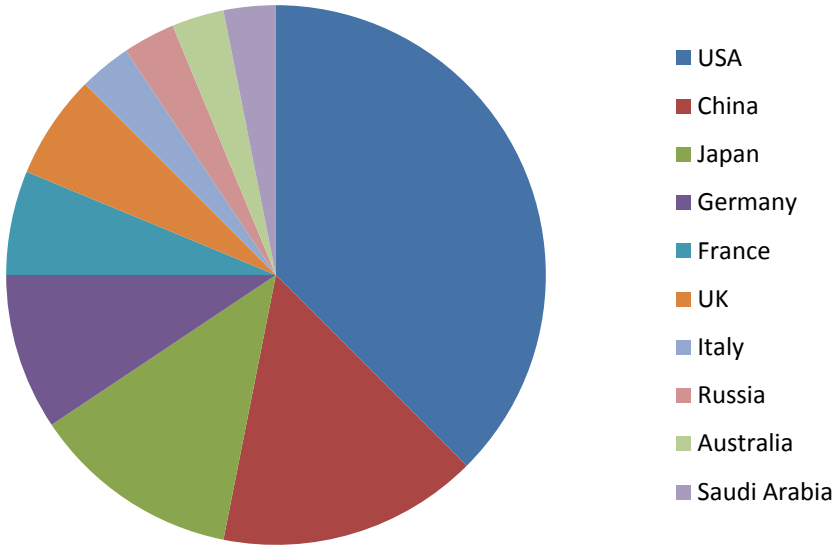

Figure 2: Petascale systems by country

The USA uses mainly the traditional architecture (50%) and accelerated (33%), with only 17% of its systems using the newer lightweight pattern. China, on the other hand, has no petascale system using the traditional architecture, and is almost exclusively dedicated to accelerated systems (80%).

Peak performance

As per the definition of this list of petascale systems, all have a peak performance of at least 1 PFlop/s. The maximum peak performance, 27.11 PFlop/s, corresponds to Titan, while the cut-off value of 1.01 PFlop/s is marked by Mole-8.5. The average peak performance in the set is almost 4 PFlop/s, yet the median lies at only 1.5 PFlop/s. This means that half of the computers are in the 1.0-1.5 PFlop/s range (very close to the minimum), yet the high performers are pulling up the average (indeed, the top 4 systems are at least an order of magnitude above the minimum).

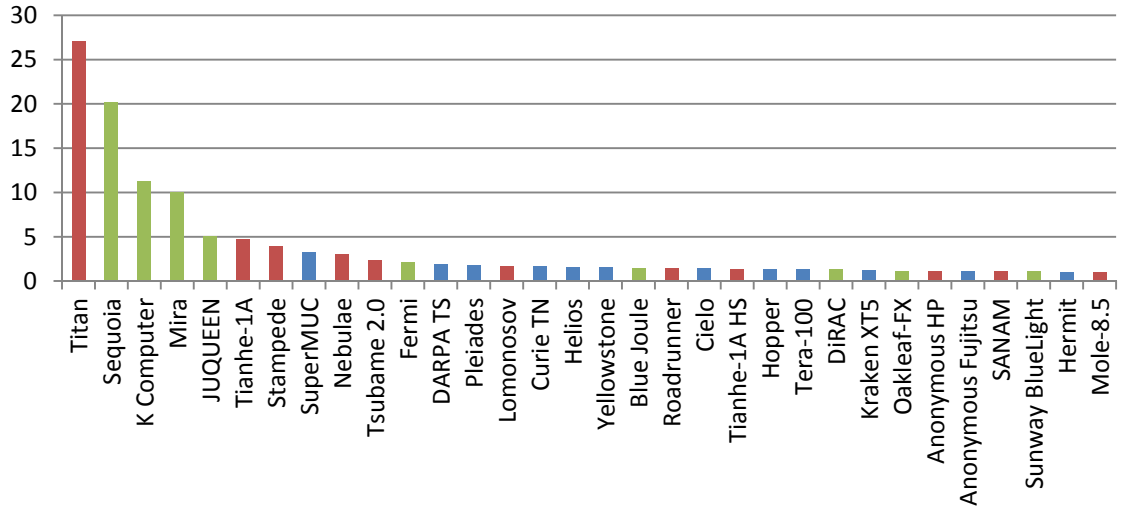

Figure 3: Peak performance of petascale systems (in PFlop/s)

LINPACK performance

Actual performance as measured by the LINPACK benchmark is considerably lower than peak performance, especially in some cases. The minimum LINPACK score is less than half of minimum peak performance at just 0.46 PFlop/s (an anonymous HP system), while the

maximum reaches about two thirds of the highest peak value: 17.59 PFlop/s (Titan). As with peak performance, the spread is quite wide owing to the big differences in performance between the top machines and the rest (average performance is 2.88 PFlop/s but the median is only 1.21 PFlop/s).

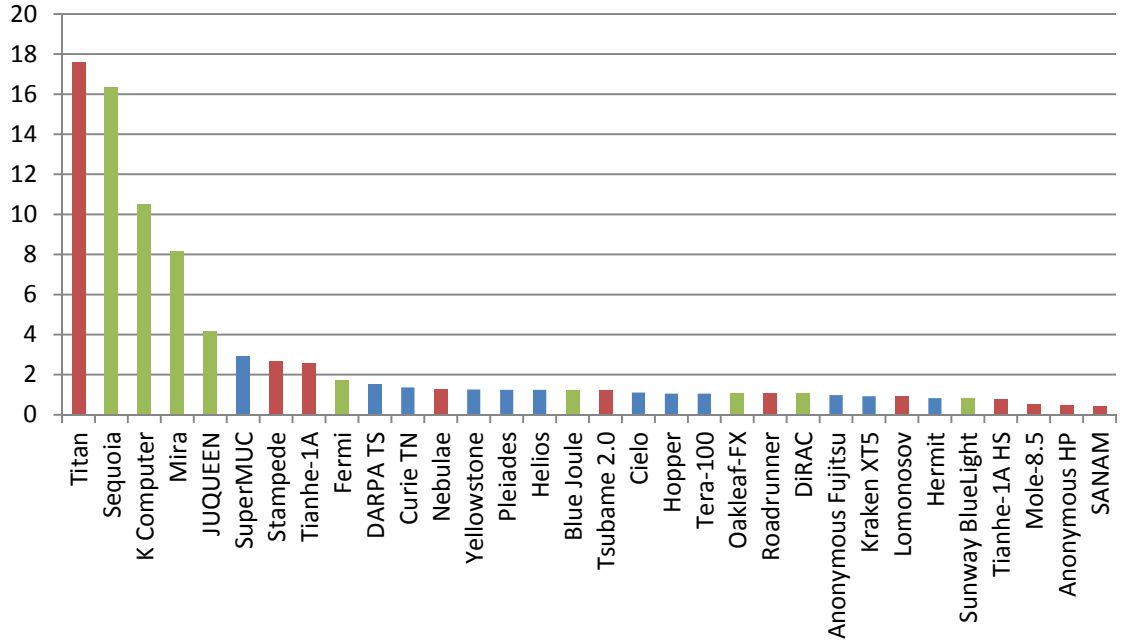

Figure 4: LINPACK performance of petascale systems (in PFlop/s)

Vendor

Almost one third of all the petascale systems (10 of 32) were built by IBM, while the next most represented vendor, Cray, only manages half that amount (5 systems that provide 15.6% market share). Bull and Fujitsu each have 9.4% of the systems on the list, which together with Cray completes the second third of market share. The last third of the market is almost uniformly shared between ten smaller vendors, two of which are Chinese non-commercial institutions: NUDT (National University of Defence Technology) and NRCPCET (National Research Centre of Parallel Computer Engineering and Technology). The commercial vendors are: SGI, NEC (whose sole machine was built together with HP), T-Platforms, Dawning, Tyan (participating in a joint venture with the Chinese Institute of Process Engineering), Dell, HP (together with NEC on one system, and exclusively on another), and Adtech. Architecture-wise, IBM is concentrated on its lightweight approach (60% of its systems), with less emphasis on their traditional architectures (only 30%), which in the case of Cray accounts for 80% of their petascale machines.

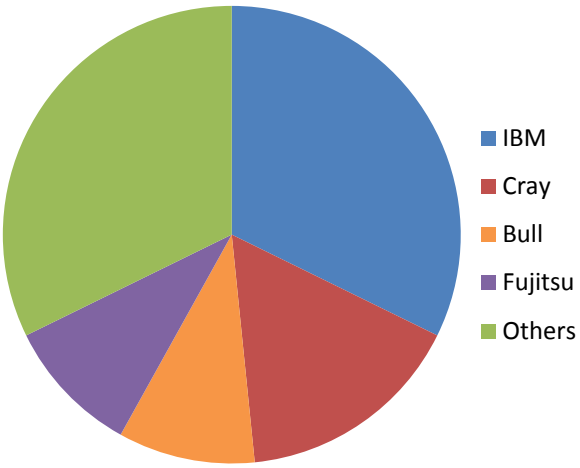

Figure 5: Petascale systems by vendor

IBM's lead is in great part due to their Blue Gene/Q model, used by 6 systems to achieve petascale (almost 20% of the total) while the next most popular model, the Cray XE6, is used in just 3 systems (10% share). The only other models that are repeated in the list are the Bull bullx B510, the IBM iDataPlex DX360M4, and the HP ProLiant SL390s G7.

Processor

Intel dominates the processor market in general, and high-end HPC is not an exception: versions of the Intel Xeon processor are found in 50% of the petascale systems. AMD, the usual runner-up behind Intel, is now tied with IBM in the fight for the second most popular processor in petascale computing, with both the AMD Opteron processors and the IBM PowerPCs (available only on Blue Gene/Q systems) taking 18.75% market share each. The remainder of the market consists of the Fujitsu SPARC64, the ShenWei SW1600, and the IBM POWER7.

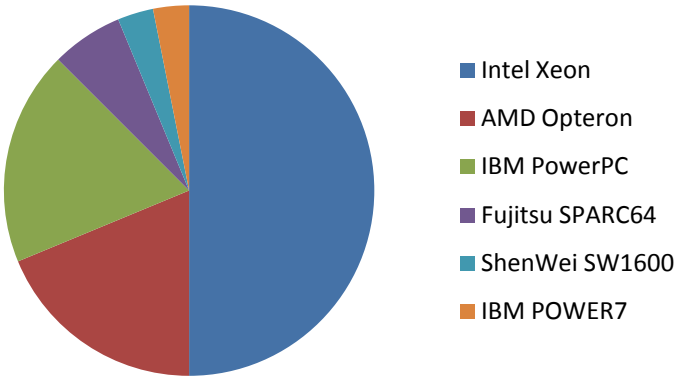

Figure 6: Petascale systems by processor

Processor clock frequency ranges from 975 MHz (ShenWei SW1600) to 3.84 GHz (IBM POWER7), though the vast majority of systems use processors with frequencies between 1.6 GHz and 2.9 GHz (90% fall in this range). The average clock speed for all petascale systems is around 2.4 GHz.

Accelerator

The accelerator market is fairly small, taking into account that almost two thirds of the petascale systems don't make use of any such co-processor. Of the ten systems that do have an accelerator, seven use some form of NVIDIA Tesla GPGPU. The remaining three co-processors found on the list are the IBM PowerXCell, the Intel Xeon Phi, and the AMD FirePro, each seen only once. It is not clear whether this market is growing (more on this in

the Dynamic Analysis and Business Analysis chapters), but NVIDIA is definitely the leader at the moment.

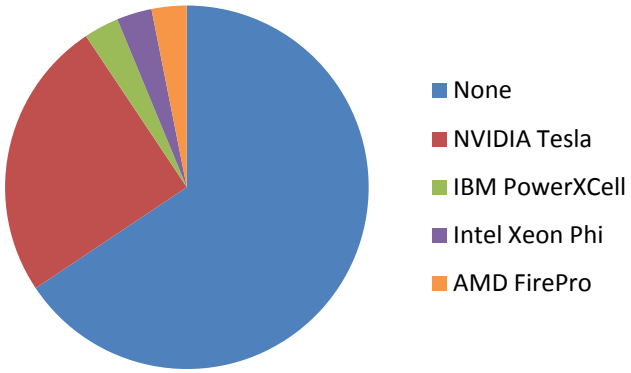

Figure 7: Petascale systems by accelerator

CPU cores

Core count ranges from around 5k cores in the case of SANAM, to more than 1.5M in Sequoia. The large discrepancy in the number of cores in these two systems, although partly due to their difference in performance, is a clear demonstration of the two main tracks taken at the moment to reach petascale performance at low power: using accelerators (SANAM) or low-power many-core processors (Sequoia).

Analysing each architecture separately we see that accelerated systems have between 4.8k and 299k cores, with an average of 62k and a median of 29k; many-core systems have between 77k and 1.5M cores, with an average of 450k and a median of 164k; and traditional systems have between 53k and 153k, with an average of 106k and a median of 113k.

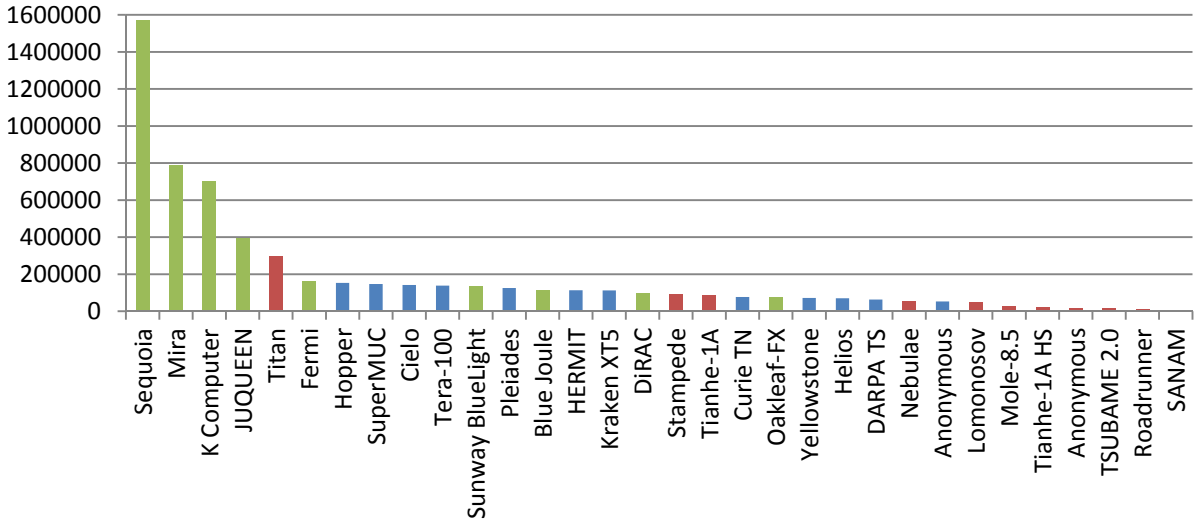

Figure 8: Core count of petascale systems

Memory

The amount of memory in each system varies up to two orders of magnitude, from 15 TB to 1.5 PB, somewhat correlated to performance but not entirely (for example, performance differences in the range are only of one order of magnitude). The average memory is around 300 TB and the median is at 207 TB. From the point of view of the architecture, accelerated systems are more “memory efficient” in terms of Flop/s per byte, and therefore have less memory than the other types of architecture for a given performance.

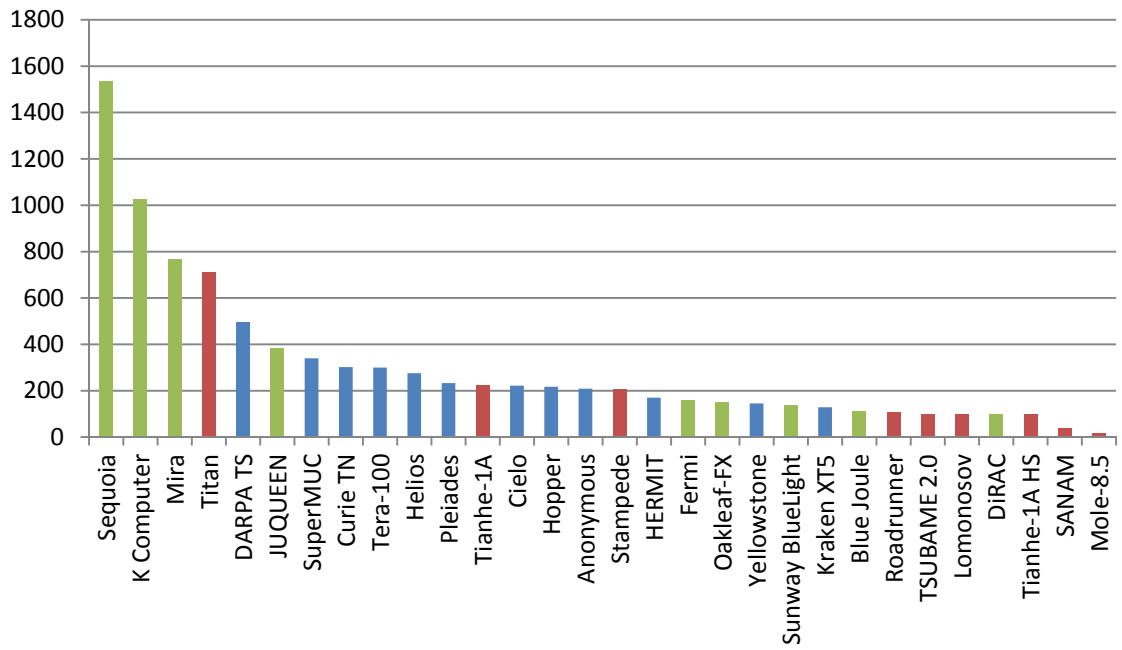

Figure 9: Memory of petascale systems (in TB)

Interconnect

A total of ten different interconnect technologies are used throughout the petascale systems, of which the main seven are:

- **InfiniBand DDR / QDR / FDR** – These three interconnects represent the successive industry standards defined by the InfiniBand Trade Association. Double data rate (DDR) has a signalling rate of 5 Gbit/s, which effectively provides 4 Gbit/s per link. Quad data rate (QDR) has a signalling rate of 10 Gbit/s, which effectively provides 8 Gbit/s per link. Fourteen data rate (FDR) has a signalling rate of 14 Gbit/s, which effectively provides 13.64 Gbit/s per link. Implementers can aggregate links in units of 4 or 12.
- **IBM BG/Q IC** – The PowerPC A2 chips in Blue Gene/Q systems integrate logic for chip-to-chip communications in a 5D torus configuration, with 2GB/s chip-to-chip links.
- **Intel Gemini** – Originally developed by Cray, the Gemini chip is linked to two pairs of Opteron processors using HyperTransport 3, and provides 48 ports that have an aggregate bandwidth of 168 GB/s.
- **Fujitsu Tofu** – Used in Fujitsu SPARC64 clusters, it is made up of 10 links for inter-node connection with 10 GB/s per link, totalling 100 GB/s bandwidth organised in a 6D torus.
- **NUDT Arch** – The switch at the heart of Arch has a bi-directional bandwidth of 160 Gbit/s, latency for a node hop of 1.57 microseconds, and an aggregate bandwidth of more than 61Tbit/s.

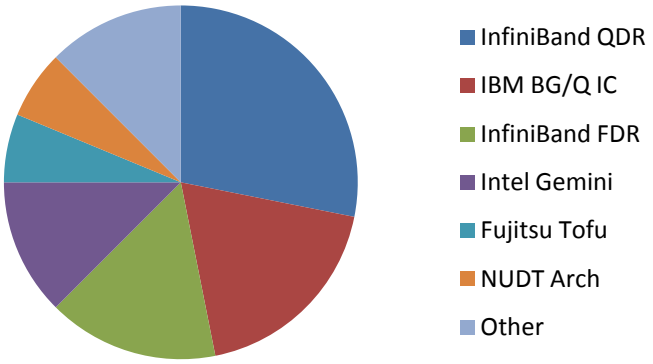

Figure 10: Petascale systems by interconnect

The most popular of these interconnects is InfiniBand QDR, with 28% of the petascale systems using it. IBM's BG/Q interconnect is used solely on Blue Gene/Q machines, yet still takes almost 19% of the market. InfiniBand FDR is now at around 16% of market share, followed by Intel Gemini (12.5% share), and Fujitsu Tofu and NUDT Arch (6.25% each). The other interconnects used are the IBM P7 IC (used in the IBM DARPA prototype), Intel SeaStar2+ (originally from Cray), Gigabit Ethernet (used by the anonymous HP cluster), and the slower InfiniBand DDR standard (still in use in parts of Roadrunner and Kraken XT5).

Computing efficiency

We understand computing efficiency as the ratio between sustained performance (executing the LINPACK benchmark) and theoretical peak performance. The value of this ratio in petascale systems is between 38% and 93%, with an average of around 73%. Similarly to core count, computing efficiency is very different depending on the architecture of the system. Accelerated systems average only 54% efficiency, with a maximum of 76% (very close to the average for all the systems). Many-core and traditional set-ups are much more similar in terms of efficiency, with many-core slightly ahead (83.4% efficiency on average, compared to 81.6% for traditional machines).

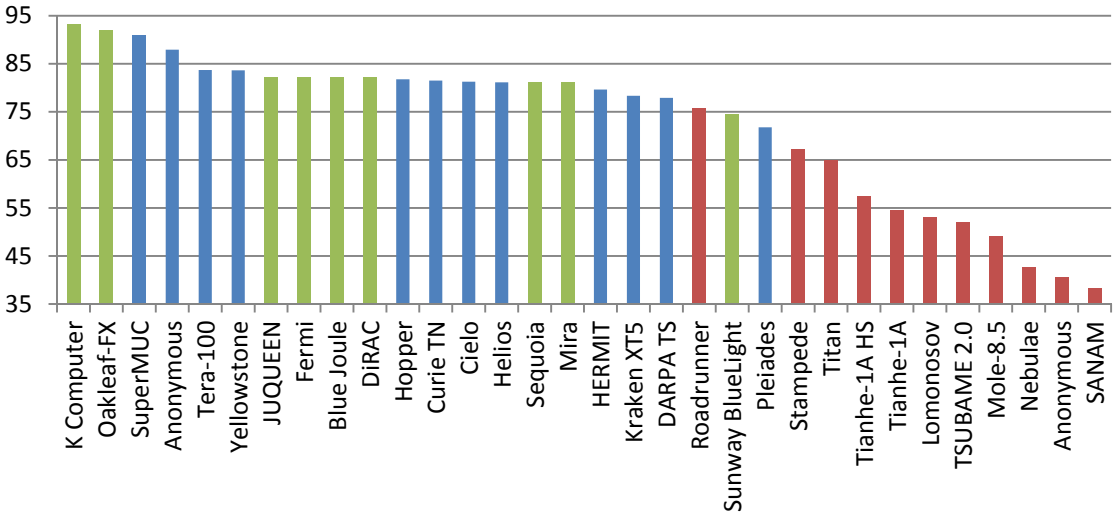

Figure 11: Computing efficiency of petascale systems (in %)

Power efficiency

One of the most important metrics today in HPC computing is the power efficiency, measured as the ratio between sustained performance (executing the LINPACK benchmark) and the

power consumption during the execution. This value, expressed in MFlops/s/W, is used to generate the Green500 list of energy-efficient supercomputers.

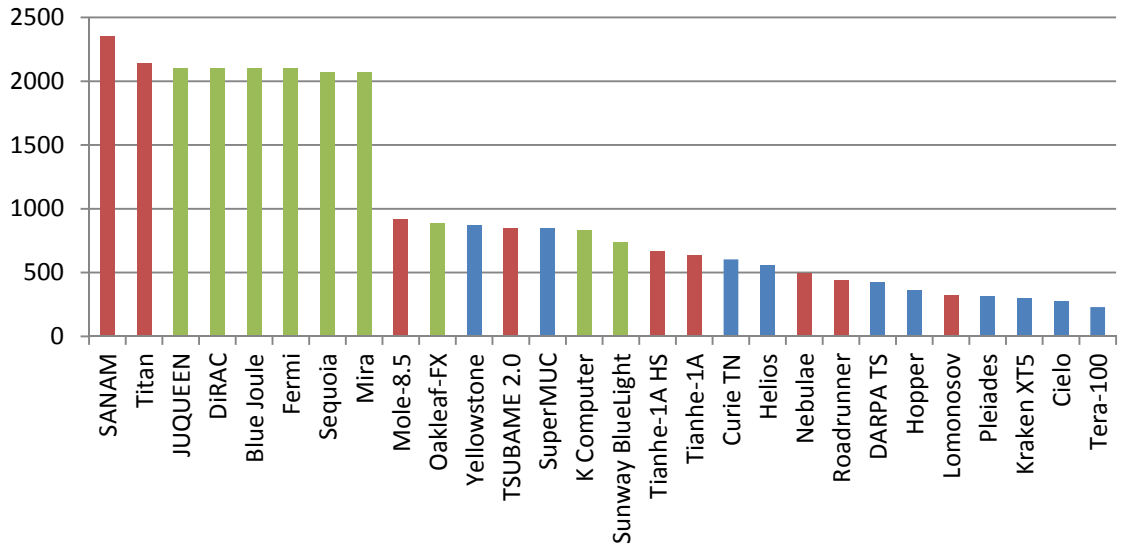

Figure 12: Power efficiency of petascale systems (in MFlop/s/W)

In this case there is an enormous difference between the eight most efficient machines (SANAM, Titan, and the six IBM Blue Gene/Q), all above 2 GFlop/s/W, and the rest, all below 1 GFlop/s/W. The overall average for all petascale systems is 1.2 GFlop/s/W, which can be decomposed by architecture as: 980 MFlop/s/W for accelerated systems, 1,667 MFlop/s/W for many-core systems, and 479 MFlop/s/W for traditional systems. This shows that the newer accelerated and lightweight architectures are much more power efficient than the traditional systems based exclusively on standard high-performance processors, with double and triple the average efficiency, respectively.

2.1.4 *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. 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.

Number of petascale systems

The number of petascale systems in the world has been practically doubling each year for the past 4 years. At this rate there will be more than 100 petascale systems in 2014, and all the supercomputers in the Top500 list will be petascale by 2016.

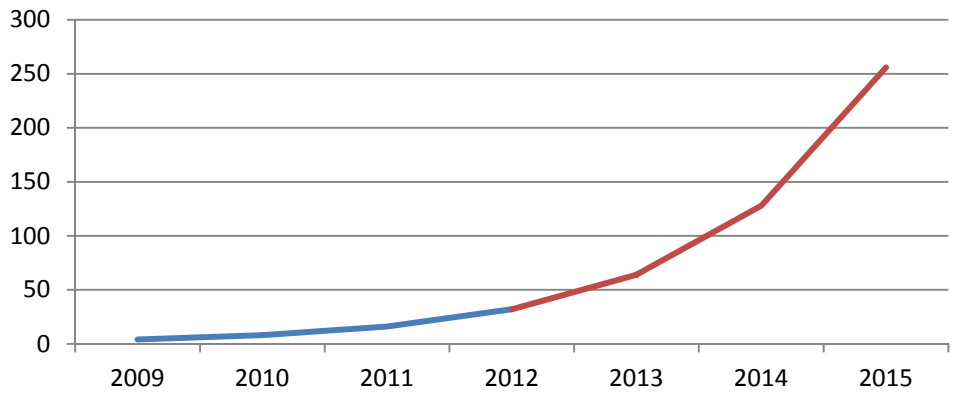

Figure 13: Evolution of the number of petascale systems (prediction in red)

Year of construction

As we have already seen, the number of petascale systems doubles annually, which means new systems take 50% of the market every year. The other 50% is distributed between the older machines (who therefore have less share each cycle), with the oldest of them slowly disappearing (in part because the share percentage approaches zero, and in part because the systems are retired and/or updated).

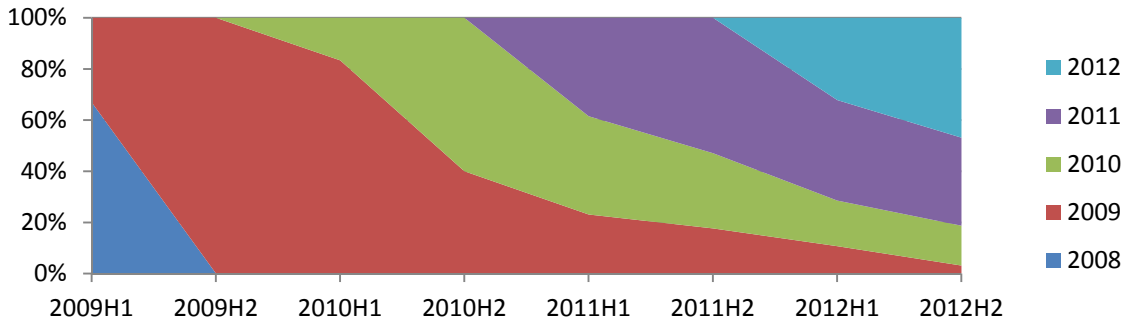

Figure 14: Evolution of the market share for construction year of petascale systems

Country

When we analyse the evolution of petascale systems according to their country, we get a glimpse not only on the geographical locations of the most powerful supercomputers, but also a slight perspective on political agendas, economic cycles, etc.

Historically, the USA has always been the leader of the top-level HPC market, with Japan as their main competitor and Europe in third place (mostly Germany, UK, and France). This has changed in recent years, reflecting a change in some countries' position and aspirations.

In 2004 China made it to the Top10 for the first time in history, by 2009 they were in the Top5, and in 2010 they took the first spot on the Top500 list. Today China is the USA's top competitor in high-end HPC, with the second-largest share of petascale computers (15.6%, compared to USA's 37.5%). At first it seemed like China could perhaps overtake the USA in the petascale niche, although the chances now appear slimmer.

Japan entered the petascale race two years late, but has tried to make a strong case for itself despite the added difficulty of competing with China as well as the USA. They have now managed to take third place with a 12.5% market share of petascale systems, but it does not seem likely that they will be able to return to second place any time soon.

Germany was the second country to have a petascale system on the Top500 list in 2009, but has struggled to keep such a dominant position. China overtook them almost effortlessly with

their first round of petascale systems in 2010, and Japan has been neck-to-neck with Germany since 2011. France joined the petascale race in 2010, while the UK and Italy have entered more recently. Together, the 4 European nations add 25% market share, which would in fact be second place between China and the USA.

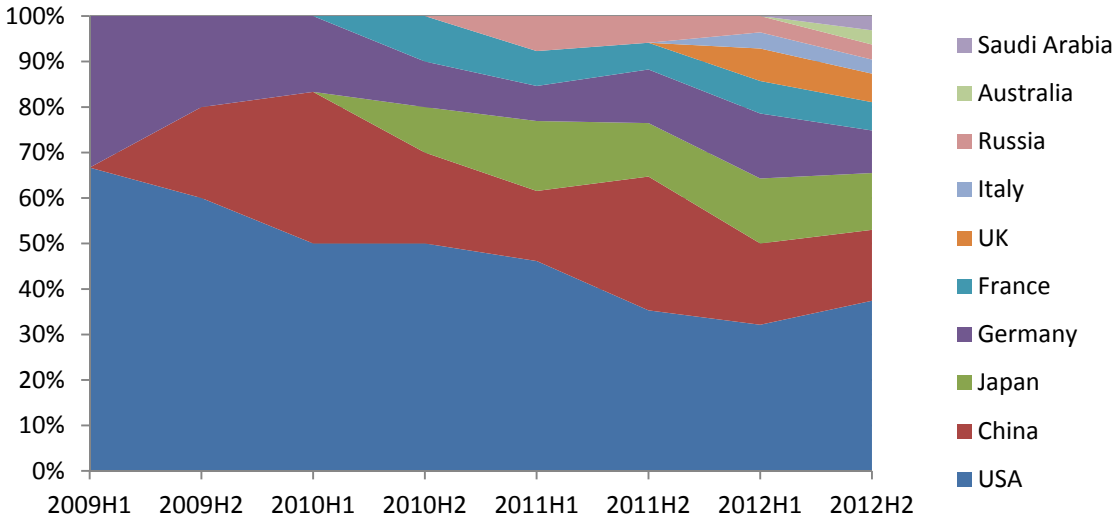

Figure 15: Evolution of the country of petascale systems

The other players are less common in this high-end HPC market: Russia (with one petascale system since 2011), Australia, and Saudi Arabia.

Peak performance

The current rate of growth for maximum peak performance is above 2.3 times annually, which means that theoretical computational power is more than doubling each year. If this trend continues, 100 PFlop/s systems should be available as soon as 2014, and the first exascale machine will appear sometime between 2016 and 2017. This might seem premature considering that estimates for exascale are usually set around 2018, but it must be noted that we are referring exclusively to peak performance.

New techniques, such as the use of accelerators to improve power efficiency, have lowered typical computing efficiency, and therefore require more peak performance to achieve similar results to traditional architectures. With a computing efficiency like Titan's (65%), EFlop/s in LINPACK would require more than 1.5 EFlop/s peak, which this model places in the 2017-2018 timeframe.

LINPACK Performance

According to this petascale-based model, LINPACK performance is actually outpacing peak performance, with a growth rate of 2.5x per year, so the lines on the chart seem to converge in 2016, which we know is physically impossible (no system can run LINPACK faster than their theoretical peak performance). The reason for this is that this model takes maximum peak performance and maximum LINPACK performance independently, which means that sometimes the values don't correspond to the same machine, therefore improving "virtual" efficiencies (the calculated efficiency doesn't necessarily correspond to a real machine) and even allowing values higher than 100%.

If it is true that in 2016 maximum LINPACK performance surpasses maximum peak, it simply means that LINPACK performance for the number 1 system in the Top500 is better than that of other systems which are bigger in terms of peak performance (but have much lower efficiencies). It is therefore an indication of a trend towards a division between high-

peak low-efficiency systems (accelerated systems) and low-peak high-efficiency systems (many-core and traditional architectures).

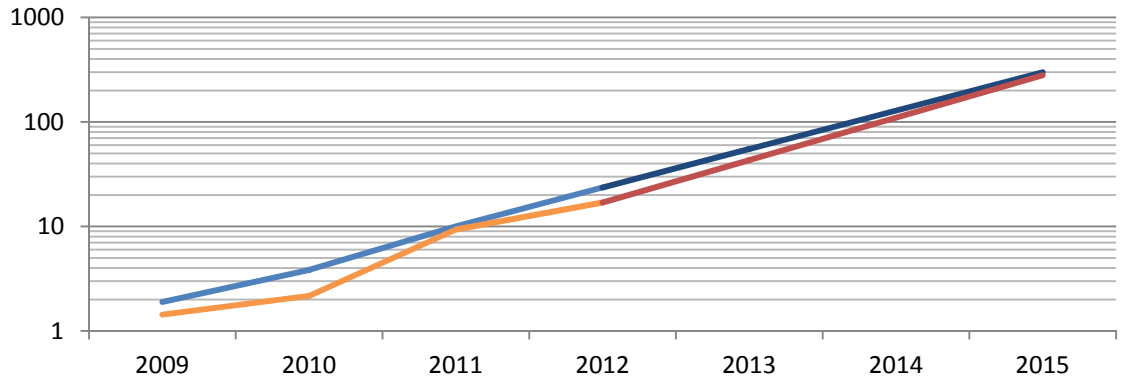

Figure 16: Evolution of maximum LINPACK (orange) and peak (blue) performance (with predictions in a darker tone)

Vendor

Petascale computing arrived thanks to the two best known HPC vendors in history: Cray and IBM. Although they still lead the market today, the road has been bumpy and their combined shares now don't reach 50%.

IBM seemed to be heading for doom in 2011, when their share had fallen to only 12% of the petascale market and the Blue Waters project was cancelled. Then in 2012 they presented six petascale systems based on their Blue Gene/Q and made a complete comeback, taking back almost one third of the market.

Cray's market share has been much more stable (thanks to their continuous introduction of new platforms: XT5, XE6, and XK7), but losing ground little by little to smaller vendors and recently also to IBM.

Bull and Fujitsu presented their first petascale machines in 2010 and 2011, respectively. Since then they have managed to retain a more or less constant share by adding a few new systems periodically. Since NUDT is not a commercial vendor but an experimental institution, it is logical that they are not striving to keep any market share. They created a revolutionary system (Tianhe-1A) to take first place, and have been losing presence ever since.

Many other vendors are starting to enter the petascale business, which is limiting the growth of the main players. It is impossible to tell how many of these new adversaries expect to grow and possibly challenge Fujitsu and Bull (or even Cray and IBM), but it is clear that the heterogeneity of the petascale landscape is giving opportunities for smaller companies to flourish while the big enterprises try to maintain their ground.

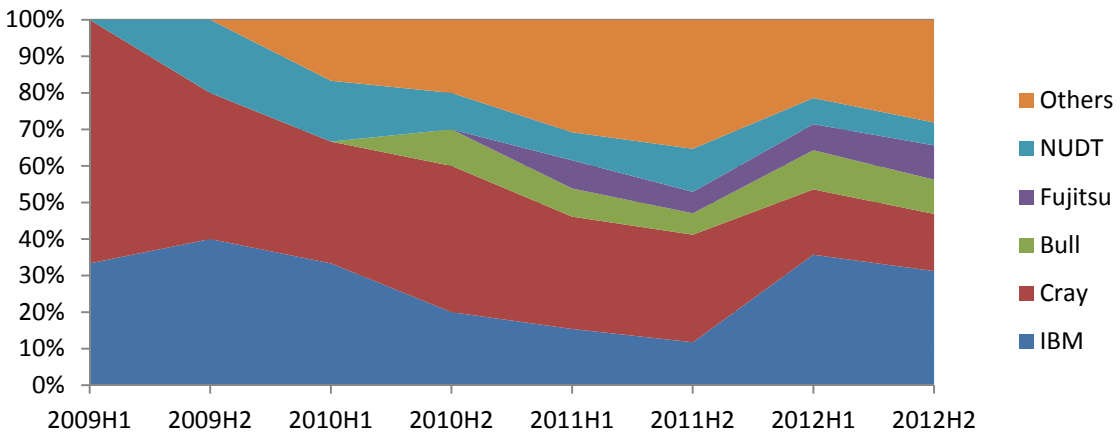

Figure 17: Evolution of vendors of petascale systems

Processor

It is interesting to see in the distribution of processors that Intel, the overwhelmingly dominant 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. 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. IBM was much more effective at this than AMD and, with the introduction of their PowerPC-based Blue Gene/Q in 2012, jumped 20% in market share (mostly lost by AMD and, slightly less, Intel). Now IBM also has a POWER7-based petascale system, which might help them add a little more to their market share in the future.

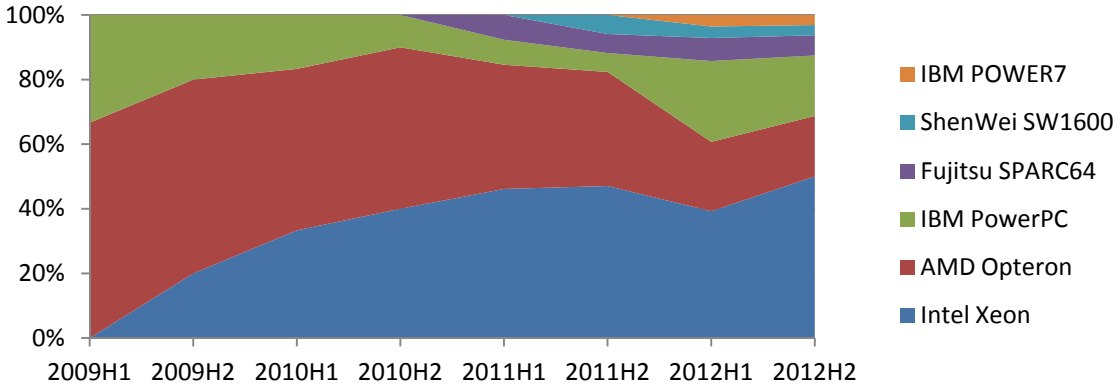

Figure 18: Evolution of processors used in petascale systems

The most surprising circumstance is the appearance, in 2011, 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). We will have to wait and see how these recent processors evolve, and whether any new chip manufacturers will join them.

Accelerators

The introduction of accelerators paved the way for petascale computing with Roadrunner, but hasn't yet consolidated a majority in the market. In fact, based on this data on petascale systems, the trend is practically flat at around 38% accelerator usage, so it is not clear whether accelerated petscale systems will ever be the norm.

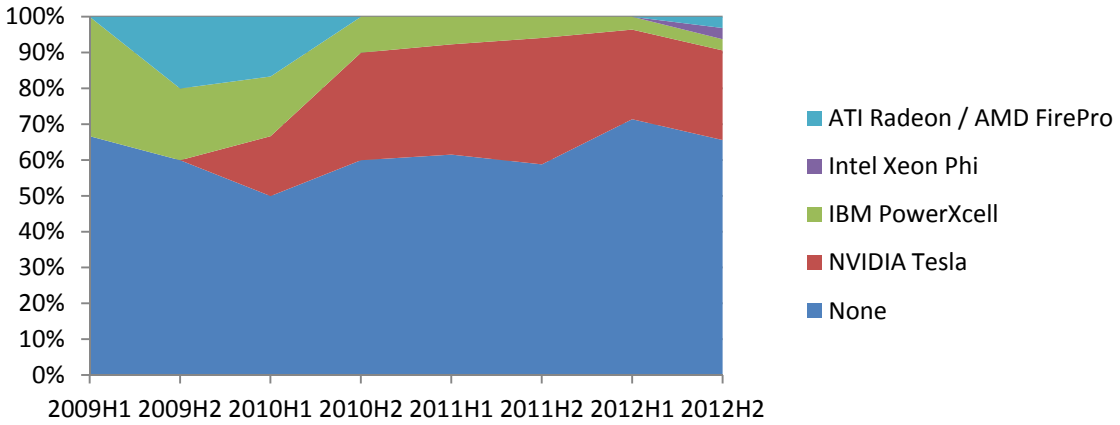

Figure 19: Evolution of accelerators used in petascale systems

The first accelerator used to power a petascale system was IBM's PowerXCell 8i, based on the Cell chip they co-developed with Sony for use in the Playstation 3 game console. At that moment the consumer accelerator market was controlled by NVIDIA and ATI, but the use of GPUs for general purpose computing (known as GPGPU) was still in its infancy. The first petascale system based on GPGPU, Tianhe-1, was launched in 2009 with ATI Radeon graphics cards. At the same time, NVIDIA was announcing plans to develop GPGPU-specific devices: the NVIDIA Tesla line of co-processing cards. Since IBM had cancelled their PowerXCell project and ATI was busy merging with AMD, the NVIDIA Tesla became the standard accelerator for HPC (including Tianhe-1A, the successor of Tianhe-1), controlling up to 85% of the accelerated petascale system market.

Currently, AMD (which now includes the former ATI) has again appeared in a petascale system with their new FirePro line of professional accelerators used in SANAM. These are not as HPC-specific as NVIDIA's Tesla line, but do allow GPGPU computations and seem to be very energy efficient (SANAM is the most energy-efficient of the petascale systems).

The newcomer to the HPC accelerator market is Intel, with their Xeon Phi (previously known as Many Integrated Cores, or MIC). This co-processor, used in Stampede, is based on a traditional x86 microarchitecture with stream processing.

Interconnect

Since the first petascale systems, the three main players in the interconnect market have been the InfiniBand standard (in its DDR, QDR, and FDR variants), IBM's custom interconnects for BlueGene/P and BlueGene/Q, and Cray's solutions (SeaStar2+ and Gemini). The industry standard InfiniBand has, more or less, hovered slightly below the 50% market share threshold, thanks to the continuous updating through its three successive generations. The BlueGene IC variants on the other hand have seen their share fall constantly after its first generation BlueGene/P supercomputer model and until the introduction of the next BlueGene/Q model in 2012. Cray maintained a high market share (around 30-40%) until 2012 with their two generations of interconnect (SeaStar2+ and Gemini), but have lost share since. The other interconnects, principally Fujitsu Tofu and NUDT Arch, share the remaining 20% of the market since they entered it in 2010.

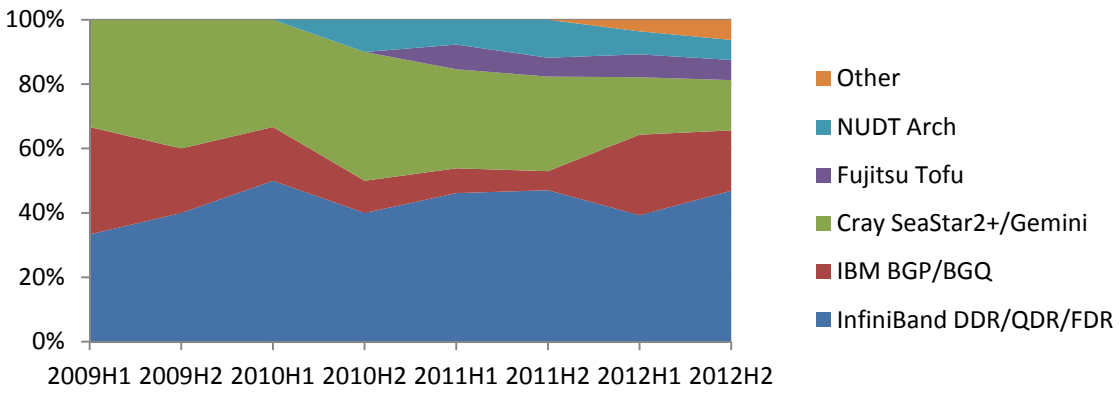

Figure 20: Evolution of interconnects used in petascale systems

LINPACK Efficiency

With regards to LINPACK execution, the efficiency of petascale systems has seen both a 10% rise in its maximum and a 37.5% decrease in its minimum. This reflects the growing difference between accelerated systems, with very low computing efficiencies and huge theoretical peaks, and many-core architectures that try to maximize efficiency of their low-performance cores. The average efficiency has been more or less constant around 71% and the median is slightly rising from 75% to 80%.

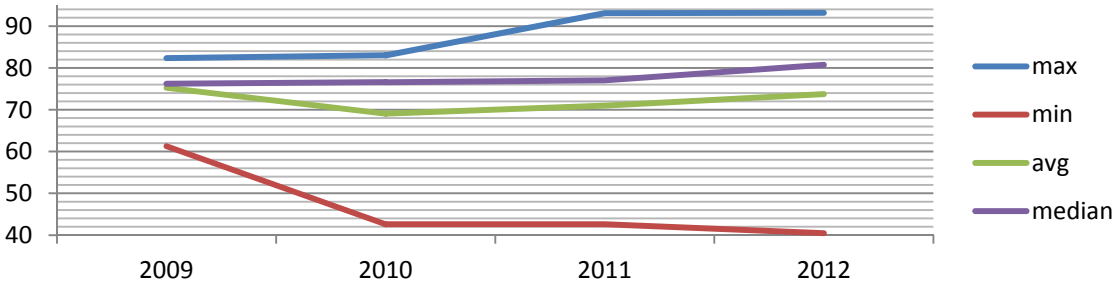

Figure 21: Evolution of the computing efficiency of petascale systems (in %)

Power efficiency

Since the power wall was identified as the main obstacle on the road to exascale, maximum power efficiency (measured in MFlop/s/W) has seen a steady growth rate of around 1.5x per year. Average and median power efficiencies of all petascale systems have also been rising by a similar amount, indicating how power-conscious the market is in general. According to this trend, reaching exascale at less than 20 MW (or 50,000 MFlop/s/W) won't be available until somewhere between 2017 and 2018, which is actually the same timeframe for exascale performance seen earlier, indicating that this trend might have been set by the inherent requirements.

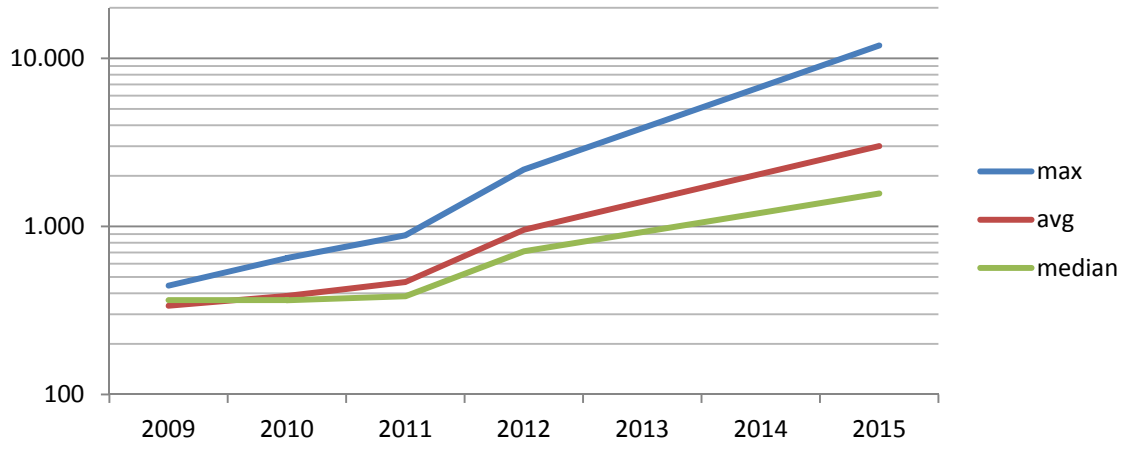

Figure 22: Evolution and prediction (from 2013 onwards) for power efficiency of petascale systems (in MFlop/s/W)

2.2 Business Analysis

This chapter tackles several topics regarding the HPC market in general from a business intelligence perspective, based on information gathered at the last Supercomputing Conference (November 2012 in Salt Lake City, USA).

2.2.1 *General HPC Trends*

Current buzzwords

For marketing purposes “Big Data” has displaced “Cloud” at the top of the buzzword list and retains the hype factor mentioned in deliverable D8.4 of PRACE 1IP. But what is Big Data? Much like the cloud it seems to be rather loosely defined at the edges.

One possible interpretation of “Big Data” is spreading compute elements among the storage. That would favour systems with interconnects capable of handling tightly coupled computations and disk I/O.

Another way to see all the marketing hype about big data is that the market expects the big growth area to be storage, and not pure computation. This would include scientific disciplines that are not in the HPC mainstream today.

Processors

While the majority of the systems in the TOP500 list are based on x86 processors, the high end of the list shows a greater variety with systems based on POWER and SPARC processors. While there are some exceptions, most non-x86 systems show up in the top 200 systems.

Intel is making a strong showing with many Sandy Bridge based systems entering the TOP500. The introduction of the Xeon Phi accelerator is also gathering a large interest.

On the x86-side AMD has kept a low profile lately. Given the company's recent ARM announcement, it seems to be hedging its bets in the CPU market.

ARM based systems regularly show up in discussions, but there is not yet any TOP500 system based on that CPU architecture. One thing holding any ARM based systems back is lack of commercially available 64-bit ARM systems. Dell for example recently showed a prototype system “Iron”, but a date for availability was not mentioned. The Mont-Blanc project continues to prototype HPC clusters using readily available ARM processors, and recently announced that Samsung Exynos 5 using both an ARM A15 core and a Mali GPU

will be used for the next iteration. In the autumn of 2012 AMD also announced that they would be building ARM based server CPUs, with production targeted at 2014.

Both the x86 and ARM ecosystems have several different vendors building both processors and systems, in contrast to the SPARC and POWER which only have systems built by the manufacturer of the CPU. This does not appear to change in the near future.

Given the increased core count per socket and power density of modern CPUs something needs to be done to meet the TDP. One way to do this what is known as dark silicon, whereby unused parts of the chip is powered off or under-clocked to conserve power. One use of this would be for co-processors only being used for specific tasks. This looks to be something that will be used more in the future, and also being used in current designs like the ARM "big.LITTLE".

Interconnect

Infiniband has the largest mind share in recent deployments, followed by Ethernet.

After Intel's purchase of the Infiniband part of Qlogic, there were some fears that Mellanox would be the sole vendor of non-integrated Infiniband adapters. Currently it seems that one can still buy Infiniband parts from Intel, but they don't appear to promote them much.

With 189 systems on the last TOP500 list having gigabit Ethernet as their interconnect, it is no surprise that several vendors (like Gnodal and Solarflare for example) are pushing for Ethernet based solutions.

Several of the top systems are using proprietary interconnects, but this fits in with the high-end being the domain of more custom-built solutions.

Storage hardware

Large disk arrays with 60+ drives in a single enclosure are now a commodity item available from many vendors.

Hard drives still hold the bulk of data, with SSDs being used for metadata and small volumes where IOPS are critical. Arrays based on SSDs, marketed as flash appliances, are therefore smaller and focused on the high IOPS market where they compete with PCI Express based products like the ones from Fusion-io.

The big data marketing push leads to vendors like Intel promoting their enterprise SSD on the same level as their CPU and accelerator products at shows like SC.

System architectures

The race towards exascale systems is producing a lot of innovation on the system side, some of which is visible even in current systems.

Several vendors are experimenting with microserver architectures on the low end, and these are architecturally showing some similarities with the IBM Bluegene/Q systems with CPU and I/O being bundled. Microserver vendor SeaMicro was acquired by AMD last year, and start-up SoC vendor Calxeda is partnering with several system vendors to try to bring ARM based microservers to market. Both these companies have focused on the interconnect fabric and a large number of smaller compute cores.

Memory bandwidth will continue to be a major issue, and will constrain the amount of memory per socket. One possibility will be to stack the memory on the CPUs, which will give a high memory bandwidth to the local cores but constrains the flexibility on memory sizes. For the medium term at least RAM will remain external to CPU socket on server class

systems. All systems can be considered as NUMA systems today, which may allow memory hierarchies to become deeper in the future.

One class of system that is not currently used in HPC is the System On Chip (SoC) based systems. These systems could enable higher growth in the microserver market, like in the embedded system market, which they dominate today. As mentioned above, these SoCs may become more feasible to HPC purposes due to an increased focus on large amounts of data. Currently they are not optimized for providing the same Flop/s/W as more traditional systems used for calculations.

A trend on the vendor side is to strive for more vertical integration. CPU vendors are buying interconnect technologies and integrating more functions into the CPU socket. This has some historical precedent, like the x86 memory controller that used to be external, but is now integrated into the CPU. With chip vendors also expanding into software, and the possibility of server platforms being SoC based, a possible future of more HPC hardware vendors being able to provide “everything” is not far off.

System software

Linux powers the vast majority of HPC systems, and this trend does not seem to subside. One notable exception is the popularity of IBM systems running AIX among the weather forecasting and climate research areas.

Microsoft Windows has a low visibility on the Top500 list with only three systems running it, and one of them is operated by Microsoft itself. Vendors advertising big data systems are more likely to use Windows, so this might be an opening for that operating system.

Many sites are using one of the enterprise Linux distributions (RHEL/SLES) or one of the community re-builds (CentOS and Scientific Linux being the most popular).

Historically it has been common for HPC sites to build their own software stack on top of a bare operating system. Several companies have tried to create a commercial product for this market, and current offerings from third party vendors like Bright Computing seems to be gaining market share.

File systems

Lustre has emerged stronger from the recent turbulent times after Oracle bought Sun Microsystems and put less emphasis on the HPC market. Currently the only real question mark regarding Lustre is the name itself, as Oracle still owns that trademark and hasn't made its intentions regarding the use of that name public.

Several vendors are selling Lustre appliances, providing integrated packages and sometimes versions of Lustre with added features. Whamcloud, the company currently coordinating Lustre development, has been bought by Intel.

A recent trend is the more direct usage of individual hard drives by distributed file systems instead of relying on RAID controllers for data integrity and redundancy. This is not new on the local file system side with ZFS and Btrfs being the most prominent examples in the Unix space.

LLNL is leading the development of a ZFS backed version of Lustre, usually referred to as the Orion branch of Lustre. Currently this is scheduled to be merged into Lustre 2.4. One side effect of this effort is a native port of ZFS for Linux.

IBM has been developing “Native RAID” for a while, and this debuted in version 3.5 of GPFS for Linux. Currently it is only supported on POWER systems and a limited number of large IBM disk systems, which are aimed at the high end of the market.

Two other cluster file systems visible at for example SC12 are FraunhoferFS (FhGFS) and Panasas PanFS.

2.2.2 *HPC vendor market analysis*

This section describes the current state of the art and planned developments of the main HPC vendors: Bull, Cray, Dell, Fujitsu, IBM, NEC, SGI, and T-Platforms.

Bull

Still improving its market share and gathering new success (Meteo France, SARA), Bull is expanding its foreign operations, which now represent half of its business.

It has two server lines, both blade-based: B500, which is air cooled, and B700, which has direct water cooling on socket ("Direct Liquid Cooling"). Both server lines can integrate Intel or AMD processors. Accelerator blades are available for both lines (B515 and B715, respectively), and can host NVIDIA Kepler or Intel Xeon PHI cards.

Bull targets HPC services and has created "Bull Extreme Factory", a Bullx based HPC infrastructure operated by Serviware (part of BULL). It targets SMEs or organization that have a temporary need for HPC resources, with access managed through a secured portal [11].

Cray

Cray has acquired Appro in November 2012, addressing a smaller-machine segment that the supercomputing company is not occupying, since their last attempt (Octigabay in 2004) was not successful. Nevertheless, their new high-end XK series is highly successful, used in Titan (> 20PFlop/s, ranked 1st in the Top500), and having a total of more than 45 PFlop/s combining all installed XK machines.

Their last high-end announcement was the XC30 supercomputer, formerly known as the Cascade program in November 2012, which is designed to support workloads of up to 100 PFlop/s. The interconnect technology is based on Aries, using a new "Dragonfly" topology. Intel technology is now integrated in XC30 blades, which is a real breakthrough as AMD was the only processor supplier on Cray's previous product line. Intel accelerator (Xeon Phi) will be supported from September 2013.

Besides, Intel has bought interconnect assets from Cray, so a question concerning possible impact of this purchase on Cray's roadmap remains.

First deliveries of XC30 have begun at the end of 2012, 3 in Europe at CSC (Finland), CSCS (Switzerland), HLRS (Germany), and 3 elsewhere: Pawsey Center (Australia), NERSC (California), and ACCMS (Japan).

Data integration is supported by SONEXION devices, a Lustre appliance from XYRATEX (European company).

Dell

Dell has a petascale-class machine ranked 7th on the Top500. Installed at TACC and equipped with Intel Xeon Phi, it reaches 2.6 PFlop/s with LINPACK. Nonetheless, their market target remains the mid and lower range of HPC clusters (10 – 100 Tflops), for SMEs and universities, aimed for Tier-1 and Tier-2 machines.

For high-end HPC solutions, Dell relies on partners: ClusterVision and Serviware.

The PowerEdge product line is composed of three configurations:

- R: Rack servers in multi-purpose classical 1, 2, and 4 based format

- M: Blade servers
- C: Micro blade servers for extreme scaling solutions

All support Intel and AMD processors, as well as accelerators.

Fujitsu

Fujitsu's high end PRIMEHPC FX10 is called the "Petascale Computer". PRIMEHPC FX10 scales up to 23.2 Pflop/s (with 98304 nodes, 1024 racks, and 6 PB of memory). The processor technology inherits from SPARC64 design IXfx generation: 40nm, 16 cores, 85 GB/s memory bandwidth, and 1.8 GHz, reaching an energy efficiency of 2 GFlop/s/W.

The interconnect, called Tofu, is a proprietary solution forming a 6D mesh/torus with 10 links (each link is 2x5GB/s), 4 RDMA engines, and 2 X 20 GB/s processor bandwidth links. This interconnect is scalable beyond 100,000 nodes (i.e. 16,000,000 SPARC64 cores).

K computer is based on PRIMEHPC FX10 technology, but using the older SPARC64 VIIIfx generation with a TOFU interconnect. Ranked #1 in the Top500 in June 2011, K computer (10.51 PFlop/s) is now at #3.

The HPC product line from Fujitu also encompasses PRIMERGY clusters, which are X86 based with InfiniBand FDR interconnect support. This PRIMERGY line is used in 7 of the 10 Fujitsu Top500 machines, two of them with over 1 PFlop/s.

IBM

Sequoia (Blue Gene/Q at DOE) has lost its first rank in the Top500 list in November, but IBM remains the undisputed leader in HPC market: 40% of all installed systems and compute power, and having 6 computers within the top 10. Their HPC software stack now encompasses IBM platform computing, although it is not yet deployed on their HPC mainstream.

Data management is targeted with innovative solutions for huge distributed data. GPFS introduces "Native Raid Server", a decentralized data processing that allows hyper scaling.

Their HPC product line is composed of:

- X series: the LRZ machine SuperMUC, ranked in November 2012, is iDataplex-based using Intel Sandy Bridge and Westmere, although the X series can also support AMD processors. GPU can be integrated and Xeon Phi will be available at the beginning of 2013.
- Power series: the first POWER7 processors are delivered in the IBM Power 775 server for HPC (best memory bandwidth per core ratio). The Flex System offers various possible configurations. Technology will move to POWER7+ (32nm), and POWER8 in 2014. Then POWER9 should be the convergence point with their Blue Gene platform.
- Blue Gene: the best seller in the HPC market, it combines high performance and power efficiency for a pretty wide spectrum of applications. Blue Gene/Q is the reference and best compromise in addressing energy efficiency and code porting. Blue Gene/Q+ should be announced for 2016. Then POWER9 may be the engine for the next Blue Gene generation.

NEC

NEC is still pursuing the vector architecture with the upcoming release of the SX-X. The processor has 4 cores with vector a performance of 256 GFlop/s, 64 GB of memory, and a bandwidth of 256 GB/s. One rack contains 64 nodes (with 1 processor per node) and provides

16 TFlop/s. The SX-X should be available in 2014, 10 times more efficient in energy than the SX-9 and five times denser.

Their software stack allows hybrid calculations with scalar X86 systems.

Pricing has been announced to be more attractive than ever. The future perspective is a completely hybrid system (vector and scalar) through both hardware and software, which should be available in 2017.

The LX Series (scalar) continues with significant successes among cars manufacturers in Europe, and recently French Tier-2 systems: University of Strasbourg and Paris-Meudon observatory compute centres.

NEC has also announced a new version of NEC LX-MPI, which is MPI-3 compliant.

SGI

SGI's HPC product line is made up of two lines:

- ICE X: Highly integrated solutions, with Intel-based blades, supporting either NVIDIA Tesla or Intel Xeon Phi accelerators. One rack provides up to 200TFlop/s, with a FDR backplane for highly scalable solutions and several topologies (Fat tree or multi-dimensional hypercube). SGI's HPC ICE X solutions remain a reference when it comes to efficiency of codes with respect to the technology peak performance and energy, as the software stack and the integration options are still at the ultimate level. Hundreds of thousands of cores can be integrated in a single system. Special software efforts are on-going to integrate and get the best efficiency from the Xeon Phi accelerators. A recent installation in Europe achieved 2.3 PFlop/s (Total, France). NASA, a historical SGI customer, has an integrated multi-generation ICE X called Pleiades with 125,980 cores providing 1.7 PFlop/s.
- UV2: Single system image, non-uniform memory access based on their proprietary internode link: NUMALink6, which keeps maximum MPI latency under two microseconds for the maximum configuration size (32,768 sockets / 262,144 cores). Multiple OS images can be deployed on a single NL6 fabric, which has a single flat address space up to 8PB and 256K threads. These capabilities allow extreme performance on a broad range of programming models and languages including OpenMP, MPI, UPC, CAF and SHMEM. Intel Xeon Phi can be integrated with the UV2 platform.

T-Platform

T-Platform is a Russian company focusing its activity on HPC. Its systems are based on the x86 architecture, integrating processors from Intel, AMD, and NVIDIA. Two keys aspects in T-Platforms systems are: high density (up to 192 nodes per rack) and energy efficiency.

The company has a global vision called "360° HPC", including technology, training and applications. Due to its size (200 people), T-Platforms has developed partnerships at various levels, including industry and academia.

T-Platforms market share is increasing in Russia, where the machine installed at the University of Moscow is their flagship model (ranked 26th in the November 2012 Top500 with 900 TFlop/s on LINPACK).

The company has first a reference approach in Europe, in Germany.

T-Platforms will launch studies in conjunction with the Russian government to develop its own processors and its own interconnect for an ambition towards exascale.

2.2.3 *HPC accelerator market analysis*

The following section presents the current state of the art in HPC accelerator technology.

Intel Xeon Phi

The "many-core" technology announced by Intel has a wide impact on the HPC market. As a matter of fact it is the answer from Intel to all GPU accelerator trends led by NVIDIA. Is it the end of an era? Is Titan the last shooting star? Energy efficiency and the ability to scale will be the key answer. The final name "Xeon Phi" enlightens the portability objective, addressing a main drawback of GPU accelerators.

What can be expected in porting an x86 code? The instruction set of "Phi" does not include MMX, SSE or AVX, it supports its own vector features (VPU) including 32 wider 512-bit registers (i7 core has sixteen 256-bit AVX registers). As a direct consequence, just listening to laziness and marketing shortcuts may lead to disappointment for performance results on Phi, as VPU has to be exploited and frequency is low. Another strong requirement is to use many threads.

The Knight should land on a socket by 2015, and become bootable natively. Without betrayal to NDA, we can expect that instruction sets from future Xeon generations should also be available on the Phi family like vector instructions and multi-core Transactional Synchronization eXtension (TSX), designed for multi core efficiency.

Energy envelope for Knights Landing should be around 15 GFlop/s/W.

NVIDIA Tesla

The new generation of NVIDIA cards, equipped with a new GPU codenamed Kepler, should be seen as an evolution of the previous generation of Fermi chips. The two most important factors are: increased computing performance with power consumption kept at the level of Fermi chips. The K20X, which is the most powerful version of the chip, doubles its predecessor's performance offering 1.31 TFlop/s of double precision performance and 3.95 TFlop/s of single precision performance while keeping the power consumption at 235W, which means 5.7 GFlop/s/W. The new cards still support only the PCI 2.0 interface.

From a programming point of view, the new architecture does not change much. There is an increased number of threads and warps, but the overall programming schema remains the same.

AMD FirePro

The new generation of AMD GPU chips introduces several major changes to the chip architecture, shifting from VLIW MIMD model to SIMD, which translates directly to better performance in computation. The GCN architecture implements several mechanisms, such as page faults to CPU OS, which help coordination of the work between CPU and GPU. AMD has also introduced support for ECC memory, which may be important from the computing resilience point of view.

The strongest card of the current generation provides 1.48 TFlop/s of double and 5.91 TFlop/s of single precision performance, consuming 370W.

In all its cards, AMD employed its "zero core" technology, which caps power consumed by an idle card to 1% of the maximum power consumption.

The AMD cards also support PCI-E 3.0 standard, which translates to double throughput (up to maximum 32GB/s) between CPU and GPU compared to the older PCI 2.0.

In the future, AMD's strategy is to merge CPU and GPU worlds together by enriching the GPU chips with the features required by C / C++ compilers to produce native code. On the other hand, in 2013 the first server APU units should be introduced, where GPU and x86 CPU cores will share a single memory interface.

FPGA

Using FPGA is not a new idea in the HPC market, but there are very few success stories so far related to employing FPGA for computing. One of the most important aspects that held it back was the difficulty of programming. In 2012 one of major manufacturers of the FPGA units, Atera, presented an OpenCL environment for its cards. This, in theory, should allow for any OpenCL enabled application to be accelerated by FPGA, which means a huge improvement in terms of the scope of supported applications. The performance and real value of these solutions still remain unknown, as there are no application benchmarks available at the moment.

2.3 PRACE and the European HPC Ecosystem in a Global Context

In February 2012 The European Commission published a communication that perfectly underlines the strategic nature of HPC [12]. This communication encompasses the whole HPC value chain from technology supply to applications through the availability of high-end computing resources (infrastructure and services) and emphasizes the importance of considering all these dimensions.

PRACE is recognized as a key player at the infrastructure level, and its sustainability considered as strategic, while the need for a European Technology Platform on the supply side is highlighted. ETP4HPC[1] is the answer of the industrial and academic European HPC technology ecosystem to this latter expectation.

PRACE

PRACE is now well established and has reached a first operational plateau with a full deployment of 6 recent petascale systems, accounting for an aggregated peak performance of ca. 15 PFlop/s, a significant fraction of which is reserved for PRACE:

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

(JUGENE, the first PRACE Tier-0 system which had been made available by GCS@FZJ in Germany, has now been decommissioned and replaced by JUQUEEN).

Although this should only been taken as an indication of the PRACE Tier-0 visibility, and not of their usage effectiveness for real applications, it can be noticed that out of these systems, 5 were registered in November 2012 Top500 list, all in the Top30, resp. at ranks 5, 6, 9, 11 and 27 – MareNostrum ranks 36 but not with its full, final configuration.

[1] The founders of ETP4HPC are the industrial companies Allinea, ARM, Bull, Caps Entreprise, Eurotech, IBM, Intel, ParTec, STMicroelectronics and Xyratex associated with the HPC research organisations BSC, CEA, CINECA, Fraunhofer, FZJ, and LRZ. ETP4HPC has already enrolled (as of writing this report) 8 new members and more are joining.

PRACE has a strong presence in both what can be called “high-power” (CURIE, SuperMUC, Hermit, MareNostrum) and “low-power” (JUQUEEN, FERMI) processor clusters. It cannot be said that one type of cluster is better or worse than the others, so having at least one of each is important so that different applications can target the architecture most fitting to its underlying algorithms. This positive diversity is further amplified by different configurations in the high-power cluster class, in term of memory per core and I/O bandwidth, which allows dispatching applications on the best suited configuration for a given project.

It is noticeable, especially by contrast with some recent US projects – Titan@ORNL, BlueWaters@NCSA but also STAMPEDE@TACC, i.e. GPGPU or manycore/MIC fuelled systems - that PRACE is still missing the same level of equipment in the hybrid cluster segment, or at least commensurable with the CPU equipment available now in Europe– only medium-size GPGPU cluster are currently available within PRACE and mostly not in the Tier-0 pool of resources made available to Regular Calls every six month.

Regarding usages, since mid-2010 PRACE has been maintaining a steady growth from 363 to the order of 1500 million core hours granted every six months through Regular Calls. Compared with INCITE in the USA (http://www.doeleadershipcomputing.org/incite-program/), a programme with many similarities which has been boosted by recent multi-petascale systems deployment, hybrid or not (e.g. MIRA, TITAN), PRACE is doing well but still lagging behind even in terms of pure (non-hybrid) CPU power.

PRACE is currently working on the definition of its Second Period, beyond 2015, time at which its Initial Period agreement will end, mostly corresponding to an upgrade or renewal cycle of the aforementioned supercomputers.

ETP4HPC (www.etp4hpc.eu)

An industry-led forum, ETP4HPC is the answer to the Commission to the need for a strong HPC technology pillar (supply side) in Europe – “competitive European HPC technologies for Europe science and industry competiveness” summarizes ETP4HPC credo and objectives.

Key European players in the field of HPC research have formed a European Technology Platform to define Europe’s research priorities to develop European technology in all the segments of the HPC solution value chain. An ETP is a well-defined kind of organization since European Commission FP6 and FP7 programmes. It is usually a not for profit association, a joint interest group of industrials with research organizations that together define research priorities in a given area, considered of importance for Europe competitiveness. There are more than 35 ETPs recognized by the Commission, out of which 9 are in the area of ICT.

The main objective of an ETP is to produce research orientations and recommendations (Strategic Research Agenda or SRA) that can then be further implemented with EC support along different operational mechanisms for R&D programmes.

ETP4HPC is currently elaborating its first strategic research agenda, planned for publication in February 2013. This is consistent with Horizon 2020 programme preparation timeline, which might fund R&D projects starting around beginning of 2014, borrowing topics and priorities from ETP4HPC SRA.

There are preliminary hints on the contents of this forthcoming SRA in a document produced by ETP4HPC in November 2012 [13]. ETP4HPC has a multidimensional vision of HPC technologies: hardware and software elements that make up HPC systems are considered first, including compute, storage and communication components, and then system software and programming environments. Then 2 axes are considered, on the one hand to push integration to its limit at extreme scale (energy efficiency, resiliency and balanced design of the system in

term of compute and I/O characteristics are critical here); on the other hand new usages of HPC are foreseen and related R&D actions proposed too (e.g. in the direction of big data handling or HPC in the cloud), as well as the expansion of HPC usages at all scales. Affordability and easy access to HPC systems, supporting the highest possible pervasiveness of HPC systems at all scales is indeed of paramount importance, in addition to exascale and beyond, since only a dense and well-articulated market at all sizes and levels of usage will ensure a lively and balanced HPC ecosystem development. ETP4HPC eventually emphasizes the importance of education and training and of the development of a strong service sector in the area of HPC, especially to accompany SMEs or larger industrial companies towards a more systematic use of HPC for their competitiveness, and proposes support actions in these domains.

PRACE (PartneRship for Advanced Computing in Europe) and ETP4HPC (European Technology Platform for High Performance Computing) are complementary and have established a constructive dialogue so that their responsibilities are clearly divided reflecting their respective domains and expertise, and that they combine their efforts to strengthen Europe’s place on the global HPC stage.

3 Hardware-software correlation

This section aims to define different correlations between hardware architectures and applications. PRACE-1IP WP8 showed that predicting the performance of applications on any given hardware architecture and vice-versa was virtually impossible due to the impossibility of characterising and classifying both hardware architectures and software applications in a useful manner. A more practical approach was to analyse the most direct interface between hardware and software, the programming models, and consider the point of view of the final users in terms of applications.

This analysis will be limited to PRACE Tier-0 systems and the applications that are run on them to guarantee value for PRACE and its members, since information on other systems can be limited and in many cases impertinent.

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 4: Programming models supported by Tier-0 systems summarizes the programming models that are supported by each of the six current PRACE Tier-0 systems, organized according to the above classification.

	Distributed Memory	Shared Memory	PGAS	Data Parallel
JUQUEEN	IBM MPI	OpenMP PThreads		
SuperMUC	OpenMPI IBM MPI Intel MPI	OpenMP Pthreads	UPC CAF Global Arrays	
Fermi	IBM MPI	OpenMP Pthreads		
Curie	Intel MPI BullXMPI MPC*	OpenMP Pthreads MPC*		CUDA OpenCL OpenACC HMPP
HERMIT	Cray MPT	OpenMP Pthreads SHMEM SMPSs OMPSs	UPC CAF Chapel	CUDA OpenACC Intel TBB
MareNostrum	OpenMPI MVAPICH2 Intel MPI IBM PE COMPS	OpenMP Pthreads SHMEM SMPSs OMPSs		OpenCL Intel TBB

** MPC [14] is a framework that implements several programming models (inc. PThreads, MPI, and OpenMP) with a number of compatibility and performance enhancements.*

Table 4: Programming models supported by Tier-0 systems

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 Tier-0 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 two of the systems (although at least two more have plans to implement some form of PGAS programming model in the near future). Data Parallel models are available on three of the six Tier-0 computers, with an ample spread of different implementations where CUDA, OpenCL, OpenACC, and Intel Thread Building Blocks are available on at least two of the systems, but never all three of them.

3.2 Applications and 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. In this sense, PRACE is well aligned with their ultimate goal of providing researchers with the maximum amount of useful computational capacity.

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.

4 Trends in HPC Energy Efficiency

The following sections provide insights on HPC energy efficiency as observed by several PRACE partners participating in WP5 during their visit at SC12. Due to the increased number of hardware components over time and as the performance continuously increases power consumption of high-end HPC systems is a serious issue that is being taken into account, discussed and researched especially towards the road to exascale.

4.1 Energy Efficient High Performance Computing Working Group

The Energy Efficient High Performance Computing Working Group (EE HPC WG) is an initiative of US organizations (LBNL with LLNL and SNL). Its objective is to drive implementation of energy conservation measures and energy efficient design in high performance computing (HPC). It answers concerns expressed by the US Federal government in the Energy Independence and Security Act of 2007 (EISA), which requires a reduction in energy intensity in all facilities, including laboratories and industrial buildings, of 30% by 2015.

The group is claiming more than 300 members from 18 different countries. The membership composition is mostly from DOE and other US governmental agencies, but also includes participants from industry, academia and international organizations, and is widely open to non-US participants. For instance, CSCS, LRZ, and CEA are members of the group from Europe.

The group is active in [15]:

- Reducing expenditure and curb environmental impact through increased energy efficiency in HPC centers;
- Encouraging the HPC community to lead in energy efficiency as they do in computing performance;
- Developing and disseminating best practices for maximizing energy efficiency in HPC facilities and equipment;
- Serving as a forum for sharing of information (peer-to-peer exchange) and collective action.

For instance, during SuperComputing (SC12 in Salt Lake City, November 2012), EE HPW WG organized a one day workshop and a BOF (Bird-of-a-feather) session.

In 2012 the EE HPW WG main activity was the definition and experimentation of a methodology for measuring, recording, and reporting the power used by a high performance computer (HPC) system - as well as other auxiliary data such as environmental conditions, while running a workload.

This is being developed as part of a collaborative effort with the Green500, the Top500, and the Green Grid. While it is intended for this methodology to be generally applicable to benchmarking a variety of workloads, the initial focus is on High Performance LINPACK (HPL), the benchmark used by the Top500. The methodology is currently being assessed and fine-tuned in real conditions by several member sites. LRZ has been taking an active part in the process, while CEA is now starting using the methodology.

Details on this methodology as well as different other documents produced by the group and other resources can be found at http://eehpcwg.lbl.gov/documents.

4.2 Energy Consumption Trends and Cooling

Systems are getting denser, but the power consumption per socket remains quite stable. For pure CPU systems a rough estimate is 100 W/socket and for GPU systems it is 200-250 W/socket. In general the power consumption is decreasing, but rather slowly. Most vendors seem to be waiting for exascale research to bear fruit before introducing radical changes.

Component level cooling is becoming more popular. Water cooling kits are now available from third-party suppliers to retrofit inside existing systems.

HPC sites are experimenting with different approaches to more power efficient cooling, such as free air cooling combined with systems that can cope with higher inlet temperatures, as well as using warm water in the cooling system. The SuperMUC is an example of a PRACE Tier-0 system using warm water, with a cooling system from the system supplier.

Which cooling solution will be the best fit varies depending on the data centre location. Conditions in southern Europe vary a lot from the ones in northern Europe, so there is no one size fits all solution.

The market for cooling without compressors is maturing with many suppliers selling equipment.

No matter what kind of cooling is used, the system outlet will contain heat. The question of what should be done with the heat that is produced remains. One popular alternative is heating the surrounding buildings, but that requires a heating system in the building that can handle the relatively low grade heat produced. For that reason it is mostly considered when constructing new buildings. Another example in northern climates is using the heat to melt snow and ice in parking spaces, which carries a relatively low cost to install.

Relocating data centres for power or cooling reasons does not seem to occur much in practice, despite talks about it. This also means that many sites have to fit systems with higher power densities into existing data centres. Encapsulating the racks is a popular solution for that issue.

4.3 Hot water cooling

One of the trends one could observe during the SC2012 conference was the introduction of liquid cooling for HPC systems. This idea is not new as in the past some Cray machines were water cooled, however due to complexity this solution was replaced by air cooling. In recent years, however, power density of HPC clusters has increased to the point where air cooling lost its major advantage – flexibility. Additionally, research done on PRACE-1IP prototypes proved that using direct water cooling may be beneficial from both the efficiency and energy consumption points of view.

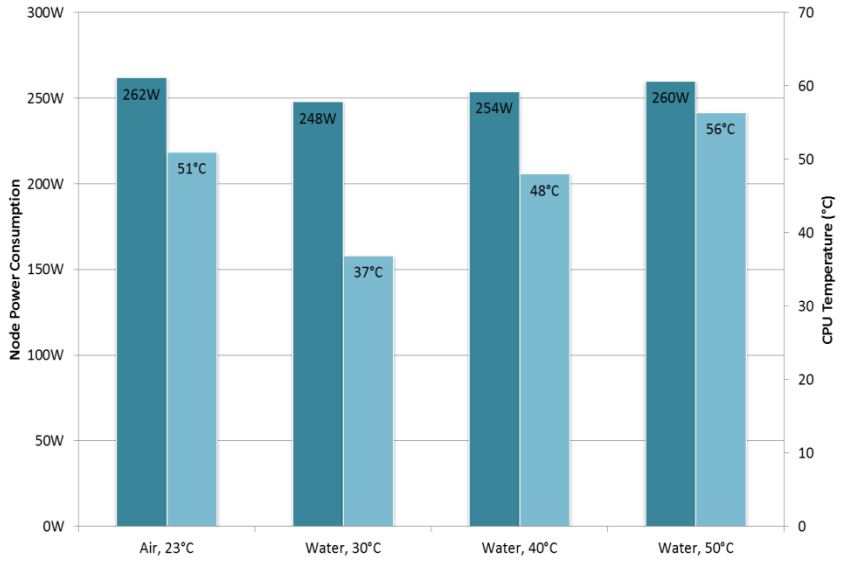

Figure 23: Dependency between coolant temperature and power consumption of the node[2].

As one can see using 40-45C water is as good from the node power consumption point of view as using 23C air, which is currently typical temperature in HPC data centres.

Using direct water cooling allows for increase of the temperature of the coolant to 50ºC or more, which means there is room for several improvements in data centre architecture. First of all, it is possible to simplify the cooling loops by removing the chillers and thus reducing energy consumption of both the computing and cooling infrastructure, therefore providing free cooling for the whole year. This subject was covered in more detail in the whitepaper on cooling technologies produces in PRACE-1IP WP8.

Some of the major HPC hardware vendors also realized the potential benefits of this solution because of either internal research or demand from customers. There are several solutions from major hardware vendors that allow for warm water cooling:

- **IBM**: iDataPlex system has a version where CPU and memory are cooled by water. There is a large installation in LRZ (SuperMUC) that is operational since the second quarter of 2012. Currently this solution allows for dealing with roughly 80-90% of energy with water because both CPU and memory are water cooled. Currently these systems cannot be equipped with GPU units.
- **Bull**: Bull took a little different approach – their water cooled servers are cooled via a large cold-plate that covers the entire motherboard, yet no information on deployed operational systems could be gathered.
- **SGI ICE**: This solution is similar in concept to IBM solution: there are heat exchangers on CPU. The memory is currently not cooled with water but on the other hand SGI have solution for cooling GPU chips.

There were also several interesting opportunities for retrofitting normal 1/2U servers with closed water loops. An example system equipped with CoolIT system was able to handle 46kW rack where 80% of the heat was transferred with the liquid cooling and disposed without use of the chillers. There are also few other solutions on the market provided by less-known companies that allow for warm-water cooling. Currently three solutions provide by Eurotec, T-Platfom and Iceotope are currently delivered for PRACE-2IP partners and will be tested within WP11.

[2] Courtesy of LRZ, Torsten Wilde and Axel Auweter

5 Best Practices for HPC Site Security

The information retrieved from the HPC centres in Europe (PRACE consortium) is based on a questionnaire including security topics. The process was started in PRACE-1IP, but because of no response from the majority of partners (former WP8 defined the threshold of above 50%) we decided to finish the process in PRACE-2IP. The process was renewed in November 2012, where we reached the demanded majority of responses. The survey allowed us to:

- Check the current state in HPC world in Europe
- Define the security level
- Present the security technology used in HPC centres
- Propose some recommendation on how to improve the security procedures.

There are many security-related technologies used in HPC centres (for further analysis please refer to the Security White Paper on the PRACE web site, when made available). The current state of implementation of the security means has been assessed based on an electronic survey initiated in PRACE-1IP WP8 and continued in PRACE-2IP WP5. Not all of the HPC centres filter the network traffic. Network firewalls are the main means of securing the networks. There should not be a single host unprotected by the firewall, especially if it can be accessed from the Internet (the acceptable exception may be hosts exchanging large volumes of data, e.g. multimedia servers). Such hosts should be appropriately placed within the network infrastructure (e.g. beyond the network firewall range) and not host any other sensitive services. Although that situation is relatively rare – it applies only to one-eighth (12.5%) of the surveyed HPCs – it is important to notice and fix the problem.

Security is a complex issue and needs to be built modularly. A single network firewall is not capable of protecting the whole network on its own. This is a reason for implementing more security mechanisms on several levels. This approach is often called “defence-in-depth”. It mitigates the security breach consequences by preventing the attacks from spreading across the network. Even if the network firewall was compromised, other security means, such as local firewalls, would stop the attack. Unfortunately almost 19% of PRACE HPCs do not use them.

Application firewalls are state-of-the-art firewalls and naturally have much more capabilities when compared to standard firewalls. Unfortunately, they have been implemented in only one quarter (25%) of the sites. This might, however, be related to the fact that inspecting all seven of the ISO/OSI layers is computationally expensive (application firewalls may also involve financial costs, e.g. for purchasing licences). One can also notice that in order to control application layer a firewall needs to know the specifics of a network flow generated by the application. Otherwise it is no better than an ordinary stateful firewall (at least not for HPC purposes).

Even the best firewalls do not fully “understand” the traffic they are inspecting. They can and should, however, be aided by host- and/or network-based Intrusion Detection and Prevention Systems (IDS/IPS). It is their task to examine the flow looking for any signs of an attack and stop it, if it occurs. Half (50%) of the surveyed centres answered they have a working network IDS/IPS, and 37.5% of them use host-based intrusion prevention/detection systems.

There are other solutions complementing HPC security, such as Data Loss/Leak Prevention systems or honeypots. They are, however, rarely used. Only one HPC has implemented a DLP system, and two HPCs maintain honeypots.

It may come as a surprise but anti-virus software in the HPC systems is not common either. Not even 19% of the surveyed centres claim to have installed such software. The common yet false belief that there are no viruses or malware in general, working on Linux-based systems

might have something to do with that result. Indeed, the greatest emphasis of the malware creators is put on Windows systems as the most popular ones, but Linux antivirus software may still be useful for e.g. protecting attached Windows-based network-attached servers or hard drives, or for email scanning. Avoiding the AV software on Linux servers as a rule may help to establish insecure practices in the organization.

More than half (56%) of the sites believe they do not need any Distributed Denial of Service protection and therefore they do not have it and are not going to implement it. Yet this kind of attack is becoming more and more popular. High-Performance Computing Centres are undoubtedly more resistant to DDoS attacks but definitely not immune. One-fifth (18.75%) of HPCs have procedures regarding combating the attacks, and the same number of HPCs have implemented an anti-DDoS software or hardware solutions. Only one centre, however, has introduced both technical solutions and procedures.

In the 13% of the surveyed centres the HPC part is located in a Demilitarized Zone (DMZ), which helps protect the internal systems and network. Three-quarters (75%) of HPCs use Virtual Local Area Networks to separate HPC infrastructure from the rest of the network but there are two centres with no separation at all. It means that a security breach on one of the machines can potentially give the attacker access to any other system located in the network, including not only servers but also network devices which often (especially older devices) have default configuration, including well known built-in passwords).

There are usually no special authentication methods used. Almost 70% of data centres use authentication basing on merely username and password. Some HPCs try other methods. One-eighth (12.5%) of them have partially implemented X.509 certificates. They are, however, issued only for a limited number of users. 6% of the surveyed organizations use other solutions basing on a public and private key pairs. One of the HPCs plans on using one-time-passwords for user authentication.

Lack of network segmentation and the use of standard password-based authentication methods are becoming even more hazardous when combined with a fact that almost 30% of the HPC systems allow remote connections via Secure Shell (ssh) or virtual desktop to privileged accounts. Fortunately, over one half of centres allow only establishing secure channel to low-privileged accounts and lets then upgrade privileges locally.

Having implemented all of the technologies mentioned above is not enough because as one of the greatest IT Security gurus, Bruce Schneier, said: “security is a process, not a product”. That is why there is a need for constantly checking and evaluating the level of security. Security audits should consist of various scenarios, such as configuration reviews and penetration tests, both black-box and white-box, from the Internet and local networks. If the HPC creates applications, code reviews are advised as well. Security audits should be performed by people who are not directly involved in infrastructure maintaining. Yet most (68.75%) HPCs have no separate team or department dealing with security issues. They, however, delegate some of their employees to it.

The survey has showed that the infrastructure security tests are usually performed periodically (at least once a year) and include mainly black-box tests from the Internet and white-box tests from the internal network. On the other hand, it is alarming that over 30% of HPCs' infrastructures haven't undergone any security test. In half (50%) of the HPC sites, security tests are performed by people administering the HPC infrastructure. It might lead to omitting certain important aspects. External tests or at least tests performed by non-involved employees are advised.

None of the surveyed centres has a separate auditing department, nor outsources it. Moreover, vast majority (88%) of them haven't undergone any formal (non-technical) security audit at

all. Also, most HPCs introduce Information Security Policies which are not based on any known standard. Only in less than 20% of HPCs the policies are based on known standards or norms, while the same number of HPC centres have no Information Security Policy at all. Finally, there is unfortunately one HPC which does not require terms of use or equivalent agreement for users to access the centre, and has introduced neither physical access control nor surveillance to protect the centre.

The White Paper describes a basic set of recommendations WP5 produces for better understanding and improving the level of security:

Recommendation 1: Perform security audits periodically

- Create or outsource a security department or team dealing with security issues.
- HPC infrastructure should undergo a security test periodically, for example every 6-12 months.
- Perform a security test of every new device or system before introducing it to the HPC infrastructure.
- Perform a security test whenever a security breach has been detected.
- If a new attack technique or critical software vulnerability has been found, perform a security audit in the involved area.
- Security tests should cover various scenarios:
    - Penetration tests from the Internet (white box and black box),
    - Penetration tests from the internal network (white box and black box),
    - Configuration reviews,
    - Application code review (if applications are developed),
- All security tests should be carried out by IT security professionals who are not directly involved in administering the infrastructure or system being audited.

Recommendation 2: Perform formal audits of the organization

- Create or outsource an auditing department capable of performing formal (non-technical) security audits.
- Create and introduce an Information Security Management System (ISMS) based on a known standard/norm (e.g. ISO27001), covering the whole organization.
- Periodically, e.g. twice a year, perform formal audits checking consistency with ISMS.

Recommendation 3: Network security

- Ensure a proper network segmentation by introducing DMZ and VLANs
- Introduce stateful network firewalls as the main network protection means.
- For better and deeper traffic inspection an introduction of the application firewalls is advised.
- Firewalls should work in the High Availability mode.
- Consider introducing a DLP system.
- Attack detection, mitigation and analysis:
    - Introduce network IDS or IPS,
    - Create honeypots for better understanding of attacks against the network,
    - Implement and introduce an anti-DDoS system.

Recommendation 4: Host security

- Introduce a local, host-based firewall on every system working in the infrastructure.
- Install antivirus software and keep it updated.
- If there is no network solution, it is recommended to install DLP software.

- If there is no network solution, it is recommended to install host-based IPS/IDS.
- Allow remote management connections only to low privilege accounts.
- Authentication should not rely only on the username and password. Introduce a higher level of security by using 2-factor authentication with, for example, X.509 certificates or one-time passwords.

The full version of the White Paper will be made available on PRACE web site at http://www.prace-ri.eu/white-papers.

6 European Workshop on HPC Centre Infrastructure

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

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).

The programmes of these workshops are well documented in D8.4 of PRACE-1IP WP8.

The Workshops Programme Committees have been composed of:

- Ladina Gilly, ETHZ-CSCS
- Herbert Huber, BAdW-LRZ
- Jean-Philippe Nominé, CEA
- François Robin, CEA
- Dominik Ulmer, ETHZ-CSCS.

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 were also PRACE-1IP WP8 or WP9 members, or are PRACE-2IP 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, and 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 now planned April 23 to 25, 2013, for two days and a half near Lugano, Switzerland, organised by CSCS (this activity will be CSCS contribution to PRACE 2IP WP5). The programme is still under construction as of writing this deliverable, but the overall foreseen structure of the agenda is the following:

- April 23 and 24 will be the plenary workshop with the usual topics (latest trends and technologies for infrastructure of supercomputing centres), likely broken down into sessions about current building project, HPC architectures and technologies, infrastructure technologies, and standards and methodologies.
  Invited talks from the Energy Efficient HPC Working Group (see section in this report), The Green Grid, ASHRAE, non-European sites like NREL (National Renewable Energy Laboratory), NERSC, LLNL, NCAR are for instance targeted, as well as a presentation of on-going PRACE 3IP pilot PCP (Pre Commercial Procurement) on Energy Efficient HPC systems, and of course a tour of CSCS new building.

- April 25 morning will be an internal PRACE HPC infrastructure session, with PRACE- 2IP WP5 (and possibly WP9) members to exchange experience and prepare PRACE-2IP deliverables.

7 Conclusion

This deliverable, the first in PRACE-2IP WP5, represents the continuation of a line of work that began with PRACE-PP WP7 and PRACE-1IP WP8. Thanks to this work, PRACE and its members now have access to the most up to date information and recommendations on the status of the petascale HPC market, interactions between hardware and software, trends in energy efficient HPC, best practices for site security, and operating HPC centres efficiently.

Future work in WP5 will be geared towards completing and formalizing this great knowledgebase, and making it more easily accessible to PRACE and its members, although much of the information has already been published in the form of white papers on the PRACE RI website or tables in the PRACE wiki.

8 Annex

8.1 HPC Centre 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 are employed?)
    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? 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
   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 the procedures)
   d. Yes, we have software/hardware solutions (please describe the solutions)
   e. Yes, we have both organizational and technical countermeasures (please describe)
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?