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 Phase Project

Grant Agreement Number: RI-283493

D5.2
Best Practices for HPC Procurement and Infrastructure

Final

Version: 1.0
Author(s): Norbert Meyer, Marcin Lawenda, PSNC
Date: 23.08.2013

Project and Deliverable Information Sheet

PRACE Project	**Project Ref. №: RI-283493**	
	Project Title: PRACE Second Implementation Phase Project	
	Project Web Site: http://www.prace-project.eu	
	Deliverable ID: < **D5.2**>	
	Deliverable Nature: Report	
	Deliverable Level: PU	**Contractual Date of Delivery:** 31/08/2013
		Actual Date of Delivery: 31/08/2013
	EC Project Officer: Leonardo Flores Añover	

Document Control Sheet

Document	**Title:** Best Practices for HPC Procurement and Infrastructure	
	ID: D5.2	
	Version: <1.0 >	**Status:** ***Final***
	Available at: http://www.prace-project.eu	
	Software Tool: Microsoft Word 2010	
	File(s): D5.2.docx	
Authorship	**Written by:**	Norbert Meyer, Marcin Lawenda, PSNC
	Contributors:	Guillermo Aguirre de Carcer, BSC Jean-Philippe Nominé, CEA François Robin, CEA Mickael Amiet, CEA Marco Sbrighi, CINECA Ioannis Liabotis, GRNET George Karagiannopoulos, GRNET Vangelis Floros, GRNET Antonis Zissimos, GRNET Vladimir Slavnic, IPB Georgi Prangov, NCSA Radek Januszewski, PSNC Lukas Dutka, Cyfronet, Mscislaw Nakonieczny, TASK Walter Lioen, SURFsara Gert Svensson, SNIC/KTH Andreas Johansson, SNIC/LiU Todor Gurov, NCSA Emanuil Atanassov, NCSA
	Reviewed by:	Thomas Eickermann, PRACE PMO&FZJ Thomas Bönisch, HLRS
	Approved by:	MB/TB

Document Status Sheet

Version	Date	Status	Comments
0.1	12/06/2013	Draft	First outline
0.2	28/06/2013	Draft	Added contributions: recommendations from the White Paper, updated contributions from HPC workshop in Lugano
0.3	12/07/2013	Draft	Added contributions: security, introduction, summary
0.4	25/07/2013	Draft	Added contributions : cooling, monitoring, energy efficiency, updated contributions: security
0.5	31/07/2013	Draft	Added: best practices in air cooling, exascalability sections, extended many subsections, text corrections
0.62	4-5/08/2013	Draft	Updated exascalability chapter, cooling, summary, general updates and compilation
0.7		Draft	Proofread
0.8		Draft	For internal review
1.0	26/08/2013	Final version	Updates and corrections done after PRACE internal review process

Document Keywords

Keywords:	PRACE, HPC, Research Infrastructure, Petascale, Exascale, security, data centre, cooling, electricity, monitoring, big data, interconnects, Top500, Green500

Table of Contents

List of Figures

List of Tables

References and Applicable Documents

[1] „SGI® Management Center,” [Online]. Available: http://www.sgi.com/products/software/smc.html.

[2] „Intel® Cloud Builders Guide to Cloud Design,” [Online]. Available: http://www.intelcloudbuilders.com/docs/Intel_Cloud_%20Builders_SGI.pdf.

[3] T. C. D. A. M. J. M. H. (. Yiannis Georgiou, „Evaluation of Monitoring and Control Features for Power Management,” [Online]. Available: http://www.schedmd.com/slurmdocs/slurm_ug_agenda.html.

[4] „Bull Extreme Computing,” [Online]. Available: http://www.bsc.es/media/4372.pdf.

[5] „Bright Cluster Manager Brochure,” [Online]. Available: http://www.brightcomputing.com/resources/Bright-Cluster-Manager-Brochure.pdf.

[6] „Bright ClusterManager 6.0 What is New,” [Online]. Available: http://www.brightcomputing.com/Bright-Cluster-Manager-6.0-What-Is-New.php.

[7] „Bright Cluster Manager 5.2,” [Online]. Available: http://support.brightcomputing.com/manuals/5.2/admin-manual.pdf.

[8] „How do I use Raritan PDUs with the Bright Cluster Manager?,” [Online]. Available: http://kb.brightcomputing.com/faq/index.php?action=artikel&cat=10&id=110&artlang=en.

[9] „HP Insight Cluster Management Utility QuickSpecs,” [Online]. Available: http://h18000.www1.hp.com/products/quickspecs/productbulletin.html#spectype=worldwide&type=html&docid=12612.

[10] „PBS Professional's Green Provisioning: Power efficiency,” [Online]. Available: http://www.pbsworks.com/Solution.aspx?lev=1&id=10.

[11] „CrayCS300-LCBrochure,” [Online]. Available: http://www.cray.com/Assets/PDF/products/cs/CrayCS300-LCBrochure.pdf.

[12] „Ganglia,” [Online]. Available: http://ganglia.sourceforge.net/.

[13] „Nagios,” [Online]. Available: http://www.nagios.org/.

[14] M. Pospieszny, „Electricity in HPC Centres,” October 2012. [Online]. Available: http://www.prace-ri.eu/IMG/pdf/hpc-centre-electricity-whitepaper.pdf.

[15] „EESI,” [Online]. Available: http://www.eesi-project.eu/pages/menu/homepage.php.

[16] „IDC,” [Online]. Available: http://www.idc.com/.

[17] „Gartner,” [Online]. Available: http://www.gartner.com/technology/home.jsp.

[18] „HPC User Forum,” [Online]. Available: http://www.hpcuserforum.com/.

[19] „Netvibes website,” [Online]. Available: http://www.netvibes.com/.

[20] „International Supercomputing Conference 2013,” [Online]. Available: http://www.isc-events.com/isc13/.

[21] J. D. Michael A. Heroux, „Toward a New Metric for Ranking High Performance Computing Systems,” [Online]. Available: http://www.netlib.org/utk/people/JackDongarra/PAPERS/HPCG-Benchmark-utk.pdf.

[22] „The Graph 500 list,” [Online]. Available: http://www.graph500.org/.

[23] „Green Graph 500,” [Online]. Available: http://green.graph500.org/.

[24] „Intel® Xeon Phi™ Product Family,” [Online]. Available: http://www.intel.com/xeonphi.

[25] „Memory System Design, ISC'13 session," [Online]. Available: http://www.isc-events.com/isc13_ap/sessiondetails.php?t=event&o=412&a=select&ra=tagcloud.

[26] „Dean Klein (Micron Technology), "New Memory Technologies to Reduce Energy", ISC'13 talk," [Online]. Available: http://www.isc-events.com/isc13_ap/presentationdetails.php?t=contribution&o=1963&a=select&ra=tagcloud.

[27] „Conclusions on 'High Performance Computing: Europe's place in a Global Race'," Brussels, 29 and 30 May 2013. [Online]. Available: http://www.consilium.europa.eu/uedocs/cms_data/docs/pressdata/en/intm/137344.pdf.

[28] „Top 500," [Online]. Available: http://www.top500.org/lists/2013/06/.

[29] „Green 500," [Online]. Available: http://www.green500.org/.

[30] „Eurora," [Online]. Available: http://www.hpc.cineca.it/news/eurora-prace-prototype-installed-cineca-new-record-energy-efficiency.

[31] „Individual ETPs," [Online]. Available: http://cordis.europa.eu/technology-platforms/individual_en.html.

[32] „Strategic Research Agenda ETP4HPC (SRA)," [Online]. Available: http://www.etp4hpc.eu/wp-content/uploads/2013/06/ETP4HPC_book_singlePage.pdf .

[33] „Working with energy aware jobs on SuperMUC," [Online]. Available: http://www.lrz.de/services/compute/supermuc/loadleveler/#energy.

[34] D. K. (. Technology), „New Memory Technologies to Reduce Energy," [Online]. Available: http://www.isc-events.com/isc13_ap/presentationdetails.php?t=contribution&o=1963&a=select&ra=tagcloud.

List of Acronyms and Abbreviations

ACE	Advanced Cluster Engine
ACF	The Advanced Computing Facility at the University of Edinburgh
ACI (OCI)	Advanced Cyberinfrastructure (ACI) USA
A/C	Air Conditioning
AICS	RIKEN Advanced Institute for Computational Science Japan
AISBL	Association Internationale Sans But Lucratif (legal form of the PRACE-RI)
AMD	Advanced Micro Devices
APC	American Power Conversion
API	Application Programming Interface
ARM	Advanced RISC Machine
ASCR	Advanced Scientific Computing Research USA
ASHRAE	American Society of Heating, Refrigerating, and Air-Conditioning Engineers
ASIC	Application-Specific Integrated Circuit
ATI	Array Technologies Incorporated (AMD)
BAdW	Bayerischen Akademie der Wissenschaften (Germany)
BCO	Benchmark Code Owner
BHO	Browser Helper Object
Blue Gene/P	an air-cooled system
Blue Gene/Q	direct water-cooled system
BSC	Barcelona Supercomputing Center (Spain)
CAF	Co-Array Fortran
CAL	Compute Abstraction Layer
CCE	Cray Compiler Environment
CCHP	Combined cooling, heat and power sytem LiU Sweden
ccNUMA	cache coherent NUMA
CCRT	The Research and Technology Computing Center
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
CMU	Cluster Management Utility
CoC	Centres of Competence
CPU	Central Processing Unit
CRAC units	Traditionally Computer Room Air Conditioners
CRAH	a computer room air handler
CRESTA	Collaborative Research into Exascale Systemware, Tools & Applications
CSC	Finnish IT Centre for Science (Finland)
CSCS	The Swiss National Supercomputing Centre (represented in PRACE by ETHZ, Switzerland)
CSTP	Council for Science and Technology Policy Japan
CUDA	Compute Unified Device Architecture (NVIDIA)
DARPA	Defense Advanced Research Projects Agency
DARPA	Defense Advanced Research Projects Agency USA
DC	Data Centre
DDoS	Distributed Denial of Service
DDR	Double Data Rate

DEC alpha	is a 64-bit reduced instruction set computer (RISC) instruction set architecture (ISA) developed by Digital Equipment Corporation (DEC), designed to replace the 32-bit VAX complex instruction set computer (CISC) ISA and its implementations
DEEP	Dynamical Exascale Entry Platform
DEISA	Distributed European Infrastructure for Supercomputing Applications. EU project by leading national HPC centres.
DIMM	Dual Inline Memory Module
DLP	Data Leak/Loss Prevention
DMA	Direct Memory Access
DMZ	Demilitarized Zone
DOD	Department of Defense USA
DOE	Department of Energy USA
DoS	Denial of Service
DP	Double Precision, usually 64-bit floating point numbers
DRAM	Dynamic Random Access memory
DSP	Digital Signal Processors
DVFS	Dynamic Voltage-Frequency Scaling
eBoost	multi-mode UPS
EC	European Community
ECCs	Error correction codes
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
ETPs	European Technology Platforms
FDR	Fourteen data rate
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
Green500	The list of most efficient systems in terms of computing power and
GTEPS	billions of traversed edges per second
GWU	George Washington University, Washington, D.C. (USA)
HA	High Availability
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.
HIDS	Host-based intrusion detection system
HMC	Hybrid Memory Cube
HMCC	HMC Consortium
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/
HPCG	High Performance Conjugate Gradient
HPCS	High Productivity Computing System (a DARPA program)
HPL	High Performance LINPACK
HT	HyperTransport channel (AMD)
HWA	HardWare accelerator
I/O	Input/Output
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)
IDS	Intrusion Detection System
IEEE	Institute of Electrical and Electronic Engineers
IEMi	adapting UPS capacity to load
IESP	International Exascale Project
Intel DCM	Intel® Data Center Manager
IPB (Serbia)	The Institute of Physics Belgrade
IPS	Intrusion Prevention System
ISC	International Supercomputing Conference; European equivalent to the US based SC0x conference. Held annually in Germany.
ISO/OSI	International Organization for Standardization/Open Systems Interconnection
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)
LAN	Local Area Network
LEED	Leadership in Energy and Environmental Design
LINPACK	Software library for Linear Algebra

LiU	The National Supercomputer Centre in Linköping, Sweden
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
MCS	Memory Channel Storage
MEXT	Ministry of education, culture, sport, science and technology Japan
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
MMU	Memory Management Unit
MOBULL	a cluster with a peak performance of 94 Teraflop/s. It is installed in a mobile data center and based on a container solution provided by Bull
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
MW	megawatts
NCAR	National Center for Atmospheric Research United States
NCAR	National Center for Atmospheric Research
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.
NERSC	National Energy Research Scientific Computing Center (USA)
NFS	Network File System
NIC	Network Interface Controller
NIDS)	Network-based intrusion detection system
NPACI	The National Partnership for Advanced Computational Infrastructure
NREL	National Renewable Energy Laboratory (USA)
NUMA	Non-Uniform Memory Access or Architecture
NV	Non-volatile (applies to memory classes or technologies)
NWSC	The NCAR-Wyoming Supercomputing Centre
Open MP	Open Multi-Processing
OpenCL	Open Computing Language
OpenGL	Open Graphic Library
ORNL	Oak Ridge National Laboratory
OS	Operating System
OTP	One-Time Password
PCIe	Peripheral Component Interconnect express, also PCI-Express
PCI-X	Peripheral Component Interconnect eXtended
PDU	Power Distribution Unit
PFlop/s	Peta (= 10^{15}) Floating point operations (usually in 64-bit, i.e. DP) per second, also PF/s
PGAS	Partitioned Global Address Space
PGI	Portland Group, Inc.
pNFS	Parallel Network File System
PRACE	Partnership for Advanced Computing in Europe; Project Acronym
PQ	Power Quality

PSNC	Poznan Supercomputing and Networking Centre (Poland)
PSU	Power Supply Unit
PUE	Power Usage Effectiveness
QDR	Quad Data Rate
RAM	Random Access Memory
RDIMM	registered DIMM
RFI	Request for Information
RFP	Request for Propsoal
RHEL	Red Hat Enterprise Linux
RISC	Reduced Instruction Set Computer
RISC	Reduce Instruction Set Computer
RJMS	The Resource and Job Management System
RPM	Revolution per Minute
SaaS	Software as a service
SARA	Stichting Academisch Rekencentrum Amsterdam (Netherlands)
SAS	Serial Attached SCSI
SATA	Serial Advanced Technology Attachment (bus)
SDK	Software Development Kit
SEC	Simple Event Correlator
SGI MC	Silicon Graphics, Inc. Management Center
SGI	Silicon Graphics, Inc.
SHMEM	Share Memory access library (Cray)
SIMD	Single Instruction Multiple Data
SLES	SuSE Linux Enterprise Server
SLURM	Simple Linux Utility for Resource Management
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
SPARC	Scalable Processor ARChitecture
SRA	Strategic Research Agenda
SSD	Solid State Disk or Drive
SSH	Secure ShellSTFC Science and Technology Facilities Council (represented in PRACE by EPSRC, United Kingdom)
SuperMUC	one of the German Tier-0 systems. Based on Intel processors with a mix of thin and fat nodes the peak performance is 3.2 Petaflop/s/s peak (Top500: #6 in the world, #2 in Europe)
SURFsara	Dutch national High Performance Computing & e-Science Support Centre (previously known as SARA)
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.
TCP	Transmission Control Protocol
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
Tier-1	Describes the second level of the HPC pyramid, usually national and regional centres

Top500	The list of 500 fastest supercomputing systems worldwide
TSV	Through-Silicon-Via
TU	Technische Universitaet (Dresden)
UCM	ultra-capacitor modules
UDP	User Datagram Protocol
UFM	Unified Fabric Manager (Voltaire)
UPC	Unified Parallel C
UPS	Uninterruptible Power Supply
UV	Ultra Violet (SGI)
VHDL	VHSIC (Very-High Speed Integrated Circuit) Hardware Description Language
VLAN	Virtual LAN
VSB-TUO	Technical University of Ostrava

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. The work package opens the possibility of closer co-operation between the PRACE community and infrastructure vendors, e.g. HPC, electricity, cooling, networking, but also security and infrastructure monitoring 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.

The deliverable D5.2 (Best Practices for HPC Procurement and Infrastructure) continues the work begun by report D5.1 published in February 2013. The additional information were compiled after the workshop organized in Lugano - 4th European HPC Centre Infrastructure Workshop (April 2013) and participation in the ISC 2013 conference and exhibition in Leipzig (June 2013). In addition, in July WP5 published a final version of the White Paper on Security in HPC Centres. This gives an overview of the current state of art of the most important security technologies used in data centres worldwide and it proposes a set of recommendations for PRACE HPC centres applicable to other HPC centres, as the white paper is publicly available. The deliverable updates the information about infrastructure issues for HPC data centres in Europe, U.S. and vendor solutions for building data centre infrastructures and smaller installations, so called mobile data centres. This illustrates important trends of the vendors' new IT architectures, cooling and electricity. A special subsection is devoted to infrastructure management and monitoring systems. A new list of the Top500 fastest HPC systems and the Green500 list of the computing systems with the best energy to computing power efficiency were released in June 2013, and D5.2 makes a

summary of petascale systems trends based on these lists and an analysis of the market. A new analysis of exascale systems and technologies required to support HPC with millions of CPUs and cores was also added: technologies for new data centres (required energy, cooling, green IT technology, the size and capacity computing rooms) and the IT technologies itself. For exascale systems, expected in 2018-2020, promising technologies and related basic research are presented.

Finally the deliverable D5.2 presents PRACE and the European HPC Ecosystem in a Global Context.

1 Introduction

The WP5 activity on Best Practices for HPC Procurement and Infrastructure is releasing the deliverable D5.2 (Best Practices for HPC Procurement and Infrastructure) with information updating those included in D5.1 report.

This new deliverable is spanning the following subjects:

- Energy efficiency in HPC
- Cooling systems and its efficiency
- Infrastructure Monitoring systems
- Power Measurement Methodology
- Assessment of petascale systems
- Hardware requirements and trends
- PRACE and the European HPC Ecosystem in a Global Context.

Additionally, the work package was also involved in the acquisition of new information that is useful in the selection of HPC computing technologies, including the determination of the power or energy efficiency of the calculations, the scalability of the solutions, new data centre technologies, exascalability and security. The collection process was possible thanks to the workshop devoted to HPC infrastructures technology (April 2013), organized by CEA (France), LRZ (Germany) and CSCS (Switzerland) – who hosted the meeting in Lugano.

Additional material was collected during the second largest yearly HPC conference, namely ISC 2013 in Leipzig (June 2013). For this purpose, the package WP5 singled out a number of issues that are within its interests. These tasks were distributed between the partners of the WP5, which also planned to attend a conference and exhibition of ISC 2013, along the following topics:

- HPC hardware requirements
- Petascacale architectures
- Assessment of petascale systems
- New CPUs
- New HPC architectures
- Exascalibility
- Energy efficient systems
- Green IT
- Solutions for data centres
- Technologies for data centres
- Cooling
- Water cooling and heat re-use
- Energy and heat monitoring systems
- Management software / operations
- Grand challenges

The deliverable is organised into 6 main chapters. In addition to this introduction (Chapter 1), the Executive Summary and to the conclusions (Chapter 8) it contains:

- Chapter 2 - Data Center buildings' ecosystem – a summary of presentations made by major sites during the 4th HPC infrastructures workshop in Lugano, including information about brand new facilities in Europe and U.S. from the academic HPC centres, national labs.

- Chapter 3 - Energy efficiency in HPC – description of the trends in cooling technologies for high density IT systems, their scalability and efficiency, electricity solutions (e.g. ultracapacitors) and infrastructure monitoring systems.
- Chapter 4 - Assessment of petascale systems – provides an analysis of what has changed between the Top500 list in November 2012 and June 2013. The chapter is also summarising the Green500 list and ideas of other benchmarks which might provide new insights on the effectiveness of HPC systems.
- Chapter 5 - Security in HPC centres – an overview of the top 8 most important security features and a summary of the conclusions and recommendations presented in the white paper produced on the topic of HPC Site Security.
- Chapter 6 – Exascalability – presents trends and basic research towards the exascale systems and data centres which will be able to support and maintain this technology.
- Chapter 7 – PRACE, ETP4HPC and the European HPC Ecosystem in a Global Context.

2 Data Center facilities ecosystem

Hosted by CSCS in Lugano, April 23-25 of 2013, also chaired and sponsored by CEA and LRZ, this workshop continued the successful series started in 2009.

60 attendees from PRACE countries and sites and from other European as well as US sites and other organizations and companies shared experience and ideas during two days of plenary sessions. This encompassed a number of recent site evolutions, technical equipment presentations, as well as a panel with HPC system vendors on paths towards exascale and how this would impact HPC infrastructures. A guided tour of CSCS new supercomputing centre ended the plenary part of the workshop.

An additional half-day PRACE closed session allowed PRACE sites to further exchange on their recent experience and projects.

2.1 Overview of HPC facitilities in Europe – Tier-0 sites

2.1.1 CEA (France)

The CEA centre located in Bruyères-le-Châtel hosts several computing centres, one for defence activities (TERA, 1.25 Petaflop/s), and one for open activities (TGCC, 2.2 Petaflop/s including the French Tier-0 system Curie). Energy efficiency is a strong concern for the two computing centres. The current PUE value for TERA computing centre is 1.35 (with is a big improvement compared to the previous value of 1.6 a few years ago). This improvement is due to better efficiency for cooling (water cooling) and for electrical distribution (less UPS) and has lead to getting the European Code of Conduct label for TERA in 2012.

An energy monitoring system provided by EDF (MAPE) was put in place at CEA. This system gathers information from 53 sensors for electricity measuring and from 27 sources for cooling measuring. This system makes it possible to measure, for example, the variation of the PUE with the computer load, with the IT room temperature and with different parameters of the cooling system. Thanks to these measures, it has been possible, with an increase of IT room temperature and a better optimisation of the cooling system, to significantly reduce the PUE in 2013 compared to 2012.

Future plans for the site in terms of energy optimization are studied in the context of collaborative projects: Cool-IT Project (2011-2012) and Perf-Cloud Project (2012-2014).

2.1.2 FZJ (Germany)

FZJ (Forschungszentrum Jülich) is a research center located in Jülich, Germany. JSC (Jülich Supercomputing Center) as part of FZJ provides services to the research community with several supercomputers. JSC hosts the most powerful Tier-0 system, JUQUEEN, a Blue Gene/Q system with a peak performance of 5.9 Petaflop/s (Top500: #7 in the world, #1 in Europe).

JUQUEEN replaces JUGENE (Blue Gene/P) since mid 2012. The transition from JUGENE to JUQUEEN was organized in several steps (during the second half of 2012) in order to avoid reducing the compute power available to the users. This transition involved the migration from an air-cooled system (BG/P) with simple heat exchangers between the racks to a direct water-cooled system (BG/Q) with, as a consequence, a stronger coupling between the infrastructure and the supercomputer.

The experience of JSC in terms of water-cooling shows that care must be taken when boards are replaced, that pressure monitoring should be designed in order to avoid confusion between leaks and bubbles and that pressure fluctuations are to be avoided. More generally, it is very important to understand the characteristics of the system and to have intensive discussions between the IT specialists and the infrastructure department.

2.1.3 LRZ (Germany)

LRZ (Leibniz-Rechenzentrum) is located in Garching. It provides a large range of services from general IT services to universities to supercomputing services to scientists from all over Europe. As such, LRZ hosts SuperMUC, one of the German Tier-0 systems. Based on Intel processors with a mix of thin and fat nodes, the peak performance is 3.2 Petaflop/s/s peak (Top500: #9 in the world, #2 in Europe). An extension of SuperMUC (phase 2, doubling the system performance) is planned in 2014.

For installing SuperMUC a new building was built in 2011 leading to a total floor space of more than 3000 m2 for computer rooms and of more than 6000 m2 for infrastructure equipment. The available power is 10 MW with redundant supply. From the very beginning energy efficiency was considered to be of paramount importance, because of the high electricity cost. Therefore a holistic approach integrating infrastructure, hardware, software and application was implemented. From LRZ point of view, preparation is of key importance when planning a new data center; this includes the need to build up expertise up front.

In LRZ it implements a novel integrated approach taking into account the energy consumption of all infrastructure components of the computing centre needed to operate the HPC systems.

SuperMUC uses mostly (≈85%) warm water cooling which, in addition to facilitate free-cooling and heat reuse, makes possible further reduction of electricity usage by removing the fans and chillers. Due to the much better cooling characteristics of water when compared to air the active CMOS components can even at water temperatures in the range of 30 °C to 45 °C be operated at lower temperatures leading to reduced leakages currents and hence an additional energy saving effect in the range of approx. 5%. Care must be taken regarding the quality of water. Some parts, including the power supplies and the IB switches, are still air-cooled; the flow of air needs to be taken into account in order to avoid hot spots. Current PUE is around 1.16, further improvement is possible but also additional investment cost needs to be taken into account.

Among the interesting results of the work done at LRZ, the work on energy-aware scheduling shows that, selecting the optimal (in terms of energy efficiency) frequency for an application, may lead to a substantial reduction of energy usage. This feature, implemented in LoadLeveler (possibly in LSF in the future) is currently used in production at LRZ, the frequency is chosen automatically based on previous runs of the applications.

Future plans at LRZ include preparing the installation of SuperMUC phase 2, improving the instrumentation, monitoring, control for energy efficiency, developing support tools and investigating further opportunities for energy re-use.

2.1.4 CINECA (Italy)

CINECA has recently upgraded its HPC platform to join the PRACE tier-0 infrastructure. The new platform is a ten rack IBM BlueGene/Q, for a total peak performance of 2.1Pflop/s.

The system is hosted in a separate computing room of about 160 m2, dedicated to water cooled systems, while the rest of the 1500 m2 data center is dedicated to other national and commercial traditional air cooled HPC equipment.

The design of the current water cooling plant originates in 2008, during the start-up of the two phase procurement process that leads to the current Blue Gene/Q platform. The total cooling power for the Blue Gene/Q room is 1.2MW (water cooling) and the control loops (as well as chillers) allows switching to indirect free cooling (two stages, two chillers). During the life of the system hosted from 2009 to 2012 in the same room, the obtained PUE (average) was about 1.45, so an additional unit for direct free cooling of the room was added in 2010. The current system/room thermodynamic behavior leads to a PUE ranging from 1.2 when the external temperature is under 9°C to a maximum of 1.5 when temperature is above 20°C, with full system load (normal production).

CINECA is investigating new hot cooled systems and especially a prototype of the EUROTECH's AURORA system (named EURORA) in order to take advantage of free cooling in the next technological step, for future procurements.

2.2 Overview of HPC facilities projects in Europe – Tier-1 sites

2.2.1 VSB-TUO (Czech Republic)

VSB-TUO (Technical University of Ostrava) hosts the only center of excellence in the HPC field in the Czech Republic, IT4Innovation.

The current supercomputer is a cluster with a peak performance of 94 Tflop/s. It is installed in a mobile data center and based on a container solution provided by Bull (MOBULL). This interim solution, in which the whole infrastructure is rented, was chosen as the final building with a new datacenter is under construction. To this supercomputer another system will be added in the future. For this extension 9 empty racks and a power of 70-90 kW are reserved. This extension will include accelerators (Xeon Phi). The target PUE for the first year of operation is 1.237; it is expected to go down to 1.208 after the extension. The experience of operating a mobile data center is positive even if setting all the equipment in less than 200 m^2 (maximum authorized by the regulation) was tricky.

The construction of the new building started in early 2013 and is on-going. It is supposed to be finished by spring 2014. This building includes a computer room of around 500 m^2. It is designed to host a large system (in the Pflop/s range) and since the system has not been selected yet it will be able to accommodate different types of systems by providing cold and hot water. Heat recuperation will be implemented for building heating and for warm water generation.

2.2.2 CINES (France)

CINES in Montpellier is the French national computing center for higher education and universities. It is currently a Tier-1 national site and its cycles (ca. 300 Tflop/s including Jade, a SGI IEC 8200 EX cluster) are pooled with those of the other national centres. CINES emploies 55 people. CINES also has a mission of data center hosting and long term preservation of documents for universities and public research organizations.

The current infrastructure has 4 rooms with a total floor space of 820 m2, with 2 primary power sources adding up to 12.6 MW. Jade compute cabinets are supplied through UPSs and have water-cooled rack back doors. Data and network cabinets, admin and connection servers are air-cooled and have two power paths by UPS and UPS plus engine generator respectively. Improvements are on-going, regarding reliability and redundancy, with the objective of reaching Uptime Tier III level, with external expertise assistance.

A new room is under construction: 600 m2 for tier-1 computers and temporary data hosting. Construction is planned for January to September 2013, then cooling and capacity components will be installed with operation starting in mid-2014. The facility is designed for flexibility, targeting direct warm water-cooling for compute cabinets (processor temperature range provisioned: 28 to 45°C), and air cooling for other equipment. Containment of data cabinets for optimized air cooling is being studied.

2.2.3 SURFsara (Netherlands)

SURFsara (previously SARA before SARA joined the SURF foundation) provides an integrated ICT research infrastructure and provides associated services. This organization is hosting and supporting the national supercomputer in the Netherlands since 1984.

The IBM system installed in 2007/2008 is currently being replaced by a Bull system. This system was selected after procurement with technical requirements based on an extensive analysis of the requirements including the interview of the top25 users of the previous system and a detailed analysis of the usage of this system. Energy and cooling efficiency were taken into account by considering the TCO of the system.

The introduction of the new system is divided into phases: phase 0 (May 2013) 89 Teraflop/s, phase 1 (June 2013) 270 Teraflop/s, phase 2 (second half of 2014) > 1 Petaflop/s. The direct liquid cooling technology is used for cooling. The inlet temperature is 30°C, allowing free cooling virtually all year round. In order to reduce the electricity consumption further, energy/application-aware scheduling technology is considered.

Regarding infrastructure, SURFsara is renting space from their commercial sister company (Vancis). The current PUE is approximately 1.5. For future systems, SURFsara is considering different solutions including a new building or renting an existing building. The expected growth of the needs for the next 10 years is expected to be a factor 2 for the power, and a factor 4/3 for the floor space.

2.2.4 PSNC (Poland)

The mission of PSNC covers different domains including HPC center, collocation and commercial data center (internet service) and network operation. The current data center (300 m2) in the city centre has several limitations: space, cooling, electricity supply, noise.

Therefore, PSNC is planning a new data-center building (1370 m2) with office space (300 staff). The location was carefully chosen, taking into account the recommendations of the PRACE white papers which are used also as best practice guides for the design of the electricity and of the cooling. The experience from PRACE prototypes is also very useful.

Meanwhile, in order to cope with immediate needs, it was necessary to put in place a container data centre (120 m2) as a temporary solution. The containers use parts as much as possible reusable in the final data centre. The containers are currently assembled in a corner on the backyard of the third phase of the future building. The transformers, the in rows air handlers and the chillers will be reused for the final DC. PSNC found that the container data centre solution, capable of 500kW, is very cheap (275 k€ including 250 k€ of reusable hardware), scalable and the process of building is especially fast.

The future building will be completed in three phases. Phase one of the new buildings will be a 1370 m2 data center. The new DC building will be composed of three levels: basement for power supply, cooling and generators plus two intermediate levels for IT rooms (one HPC floor suitable for high power density, one Internet service floor for low power density). Thr roof will be used for liquid chillers. The center is planned to scale up to 20MVA of total electrical supply (transformers) and 14MW of cooling power placed on the roof. PSNC is exploring the possibility to use the cold water of Warta River for adiabatic cooling.

2.2.5 CSCS (Switzerland)

CSCS moved into their new supercomputing centre in spring 2012 – it was publicly inaugurated in the end of August 2012. Started in 2008, this project achieved exactly the planned timeline, including an aggressive plan for moving all the existing hardware and the staff (resp. in 3 and 2 weeks).

The two-component facility (office building, technical building) is next to Lugano fire brigade – which can intervene very quickly: this allowed CSCS to elect not to install a fire extinguishing system, only detection and power cut. Specific collaboration between CSCS and the fire brigade has been developed.

The building is designed in a flexible and modular fashion in order to accommodate a life expectancy of circa 40 years, with no pillars in the machine room. The lake water cooling relies on a 2.8km pipeline and has a total capacity of 700+ l/s. A separate reservoir has been built to allow the local industrial works to also benefit from the pipeline.

Enclosed cooling islands are set up in the machine room (low and high density resp. 10 kW and 30 kW per rack). This puts some constraints on hardware that can be accommodated within them (form factor, weight, power, etc.)

Supercomputers are not on UPS. So far, this choice has worked well in spite of electric storms that can sometimes bring the equipment down. The target of a PUE of 1.25 or less has been reached.

A few issues were encountered and tackled: alignment and stability of raised floor, network problem on lake water installation, PVC in filters in front of main heat exchangers, fine tuning of pumps and building automation.

With hindsight, all design choices were proven relevant, only a few adjustments should be considered – such as lake water cooling redundancy, raised floor weight test.

2.2.6 IPB (Serbia)

A PRACE partner since 2008, Serbia is represented by IPB in the consortium. IPB gained substantial experience through participation in Grid projects (series of EGEE and SEE-GRID projects, and on-going EGI-InSPIRE project). IPB also participates in a regional HP-SEE HPC project.

The current IPB cluster PARADOX consists of 84 nodes with 2*4 core Xeon E5345 processors and 8 GB of RAM each. Procurement for upgrade was finalized in December 2012, with a two-round tender assessing price, performance, interconnect, storage, parallel file system, cooling system and energy efficiency. The solution will be a 106 TFlop/s (theoretical peak performance) CPU/GPU HP Infiniband cluster with Proliant SL250s nodes – 16 CPU cores = 2 Intel Xeon E5-2670 + 1 NVIDIA Tesla M2090, 32 GB RAM each. Racks will be water-cooled.

Storage system will provide circa 100 TB of space, with a Lustre parallel file system.

Facility upgrade is in progress to accommodate the new cluster. Its total power consumption will be 100 kW including legacy equipment, and for cooling a HiRef free cooling chiller will be used (124 kW capacity).

PARADOX will become a PRACE Tier-1 system.

2.2.7 LiU (Sweden)

The National Supercomputer Centre in Linköping, Sweden, provides HPC services to academic institutions in the country as well as to partners SMHI (meteo) and SAAB, since 1989, with a staff of 30 people. It belongs to SNIC meta-centre. HPC resources deliver ca. 500 Tflop/s and 4 PB of data storage.

Two computer rooms of resp. 2003 (120 m2, 160 kW) and 2007 (240 m2, 840 kW) are currently in operation with air-coolid and water-cooled racks. The more recent facility with mixed air/water cooling has a PUE of 1.17. A raised computer inlet temperature (20-25°C) reduces chiller time and increases the efficiency of cooling equipment as well as the possibility of energy re-use.

A new computer room building (4 cells) will be ready in summer 2013. The first cell is arranged for 1 MW of maximum computer load, a hosting capacity of 80 racks and air cooling with aisle separation. Water cooled systems as well as containerized units can be hosted in the future in the other cells of the building.

Cooling is provided by a municipality owned company in charge of district cooling and heating. This company operates a combined cooling, heat and power sytem (CCHP). A waste incineration plant and boilers powered with biomass, coal and rubber or petroleum can deliver power and district heating. Cooling is produced in winter by cooling towers and in summer by absorption chillers. At this time the output temperature from the IT equipment is too low for the heat to be reused in the district heating system. This may be the case in the future if output temperature from IT equipment is high enough (> 50°C).

2.2.8 EPCC (UK)

The Advanced Computing Facility at the University of Edinburgh hosts several large systems including a CRAY XE6 system for the UK national academic service (Hector) and a BG/Q system (DiRAC) for theoretical modelling and HPC-based research in particle physics, astronomy and cosmology. The ACF is staffed and operated by members of the Edinburgh Parallel Computing Centre (EPCC) which is part of the universities school of physics and astronomy.

The current facility has been in operation since 2004 and is located near to Edinburgh where it can take advantage of space and ample power supply. Prior to operating from the ACF, the EPCC hosted HPC systems on the Universities Kings Buildings campus. Today, the ACF includes three computer rooms (2x280m2 1x 480m2) and the present electrical capacity is 6 MW (> 4MW deliverable power/cooling).

The University is mandated to provide an operating environment as energy-efficient as possible; therefore the ACF implements several measures in order to improve the energy-efficiency. This includes free-cooling, optimised air-path with minimal mixing of supply/return air, effective regulation of chilled water-flow; AFD controlled pumps/fans and optimised chiller performance.

The free cooling system has proven to be useful all year round as data from 2011 and 2012 shows that every week of the year, some level of free cooling was achieved. Full free cooling, with chillers disabled, was in operation for roughly 20% of the year

For Archer (the next UK national service), which is planned to go live in Q3/4 2013, the ACF has been extended with a new computer room of 480 m2 and an adjacent plant room space of 760 m2. The new plant has a 4 MW cooling capacity and an authorised capacity of 6 MW from the local power company. UPS will be used for services such as disks, admin hosts and critical plant infrastructure whilst a 2MW diesel generator will back these services. In addition to the building upgrades the network connectivity with JANET has been recently upgraded by the installation of dedicated fibre for the ACF.

2.3 Overview of HPC facilities projects in Europe – Other sites

2.3.1 TU-DRESDEN (Germany)

TU-DRESDEN is operating two kinds of systems with very different requirements: first, HPC systems and second, IT equipment for the University Hospital. The first needs hot water cooling (>30°C) and also cold water cooling (<20°C) with moderate availability, while the second needs efficient high density air cooling (20kW per rack) and high availability.

TU-DRESDEN is currently building a new site for a new data-center which should be suitable for both kinds of systems. The project started in January 2012 and is planned to finish in October 2014. The global target is a PUE<1.3 with waste heat-reuse in campus.

Given the driver for the room design is an efficient, flexible and scalable high density air cooling, various solutions were studied. The preferred one is with hot aisle containment, due to the equivalence in term of efficiency with cold aisle containment but with several advantages. These advantages include:

- minimize hot air volume and leakage,
- minimize air-side pressure drop,

- minimize exposure of personnel to hot air,
- allow for high density CRAH placement.

During the design of the modular room infrastructure TU-DRESDEN realized that placing the CRAH directly under a suspended floor, in a spacious plenum, can allow for a clear management of the mechanical equipment, which is completely separated from the IT equipment placed above the floor. Upper floor is the domain of IT personnel, IT equipment and data cabling while bottom floor is the domain of infrastructure personnel and mechanical equipment. The advantages during operations are: operational reliability and maintainability, simple interfaces, spacious access, maximum flexibility for construction in a live data center, more room for power rails and piping. Day to day maintenance and operation, like pipe work that may produce dust, are facilitated due the "dirty" work occurs not on IT floor. Studying on how to reuse the waste heat, it was pointed-out that very often the design is driven by peak parameters, in terms of equipment heat production (e.g. running Linpack) and building heating requirements (lowest winter outside temperatures). The design of waste-heat reuse loops should be carefully optimized based on the average HPC load and average winter temperatures, allowing for an external heat supply to cope with peaks demand.

Few challenges remain: control of hybrid cooling towers (cost of water for adiabatic cooling vs. cost of electricity for fan power); control loop for chiller vs. free cooling; control loop for cold water temperature (increase temperature at partial load); water quality issues (ASHRAE vs. HPC vendor specifications).

2.4 Overview of HPC facilities projects in US

2.4.1 National Center for Atmospheric Research (NCAR)

The NCAR-Wyoming Supercomputing Centre (NWSC), which opened in October 2012, is a world-class centre for high-performance scientific computing in the atmospheric and related sciences. The largest system installed is Yellowstone (IBM iDataPlex) with a peak performance of 1.5 Petaflop/s. The computer centre architecture is "data-centric" with a large central disk resource of 11 PB (planned to be upgraded to 16 PB in 2014) and a HPSS archive of 100 PB. The power efficiency on NCAR workload of Yellowstone is around 43 sustained Megaflop/s per Watt.

Regarding the facility energy efficiency (operation cost saving and emissions reduction) is the main goal but risk minimization and efficient use of capital are to be taken into account.

The approach for improving the energy efficiency focuses on 4 main points:

- Utilisation of the regions cool and dry climate;
- Utilisation of liquid cooled computer solutions where practical - using evaporative cooling towers to efficiently deliver sufficient cooling capacity directly to NCAR's supercomputer for 96% of the year;
- Utilisation of hot aisle containment for commodity equipment - waste heat from the supercomputer is captured and reused to heat administration areas of the building and to melt snow and ice on exterior walkways and loading docks;
- Focus on the biggest losses (compressor based cooling and transformer losses).

Almost 92% of the NWSC energy is going directly to its core purpose as a data center: powering supercomputers to enable scientific discovery.

The current PUE, after 9 months of operation, is around 1.1. Beyond PUE, the optimisation of energy efficiency is challenging because of considerable computer load variability (for Yellowstone, 300 kW idle to 1700 kW peak demand) and of the weather variability. A host of ultra-efficient water-conserving technologies facilitate savings of up to six million gallons of water per year. In practice, real cost savings are in part dependant on natural gas price as it is currently somewhat cheaper and lower emissions to heat the facility with boiler systems rather than heat pumps. An interesting fact towards sustainability is that renewable wind energy provides direct power to the facility, starting at 10% of supply with the ability to raise that percentage as conditions permit.

For NWSC the lack of skilled engineers in HPC centre facilities is a challenging recruiting problem. Therefore, NWSC acts as a teaching laboratory and hosts students for work on practical facilitiy topics with the intent of growing expertise in the field.

2.4.2 National Energy Research Scientific Computing Center (NERSC)

NERSC is the primary computing facility for DOE Office of Science. NERSC currently operate several large systems, the largest being Hopper (NERSC-6) (CRAY XE6, 1.3 Petaflop/s peak).

NERSC will move to a new building (CRT) in fall 2014. This building, under construction, will be a mixed office (300 offices) and data centre (20000 sq.feet expendable to 28000 of HPC floor) building. It will be extremely energy efficient including free cooling for air and water, heat recovery and LEED gold design. The power capacity to the building is 42 MW with 12.5 MW at move-in. The location of the CRT was chosen in order to increase collaboration with the departments of the University of Berkeley.

The predicted PUE is 1.1. Due to the local weather conditions and location (hillside) free cooling is possible all the year for air-cooling of 75F and water-cooling of 75F. For water-cooling at 65F chillers are necessary 560 hours/year -6%).

Regarding air-cooling, the high variation in humidity may be a problem for tapes, a solution is to install the tapes in a different room. Another issue is office heating when computers are stopped, adding boilers is a solution under study.

A specific problem of NERSC is the location very close to a major earthquake fault. Therefore, seismic isolation floor will be installed in the CRT.

The move to CRT will take place in 2015 without centre shutdown. Hopper will stay in the old building until it retires in 2016. Edison (NERSC-7) will move in early 2015.

2.4.3 National Renewable Energy Laboratory (NREL)

The NREL is the U.S. Department of Energy's primary national laboratory for renewable energy and energy efficiency research and development.

The new high performance computing (HPC) data center in NREL’s Energy Systems Integration Facility (ESIF) is designed to be one of the most energy efficient data centres in the world by deploying evaporative cooling; featuring warm water liquid cooling and waste heat capture and re-use. The NREL HPC data centre is build as a showcase facility with a goal of 1.06 in terms of PUE reached by leveraging a favourable climate and waste heat captured and used to heat labs & offices.

A holistic approach is necessary for sustainability in a context where data centres are highly energy-intensive facilities. Today, in the US, data centres electricity consumption is around 3% of the total electricity consumption and this figure is projected to double in the next 5 years. Sustainable computing involves choices regarding power, packaging, cooling and energy recovery in data centres that should take into account carbon footprint, water usage and costs. A holistic approach to sustainability and TCO is needed for the entire computing enterprise, not just the HPC system.

Regarding cooling, liquid cooling has a lot of benefits including a better thermal stability for components, better waster heat re-use options, reduction of condensation and suppression of expensive and inefficient chillers. It is highly recommended to follow the latest ASHRAE water quality specifications in order to avoid troubles. Use of evaporative cooling should not be discarded since it doesn't necessarily result in a net increase in water use if you consider the water use for production of electricity.

Sustainability requires a 3-D optimisation along three axes: IT power consumption, facility PUE and energy reuse. PUE is not enough in order to assess the efficiency of a data centre, ERE, which takes into account heat reuse, will help drive sustainability. The focus needs to be know on compute efficiency and energy reuse.

2.4.4 ORNL

ORNL currently operates multiple petascale systems in the NCCS (National Center for Computational Science) facility, mainly: US DoE Titan 27 Pflop/s peak; NSF Kraken 1.03 Pflop/s peak; NOAA Gaea 1.1 Pflop/s peak. The TITAN installation was executed in two phases, the second completing in 3QFY13. Titan was measured at 9MW (peak) power consumption during HPL.

Current facilities are based on a building with 2x1860 m2 traditional 3 feet raised floor, powered by 16 MV-LV transformers, for a total of 26.5 MVA, cooled by 5 chillers capable for a total of 6600 tons at 5.5°C water (supply) with 7.3°C DeltaT. The primary UPS is rotary equipment for a total of 1000kVA with a 1500 kVA backup diesel generator. There is a secondary battery-based UPS with capacity of 500kVA with a 750kVA backup diesel generator.

Furthermore ORNL is preparing a new HPC step to be completed in 2017, for a new 20MW installation. It will be part of a big joint procurement under the CORAL umbrella (Collaboration between ORNL, ANL, LLNL) for three new pre-exascale (100 Pflop/s) systems each one for up to 20MW of total consumption.

For the CORAL system at ORNL the plans are for a new 1000 to 1200 m2 floor (not raised). The chilled water and electrical cabling will be from above. The expansion will be powered by further six transformers, 3/4MVA each. On the cooling side, additional 4500 tons of chilled water are requested. They are assessing a new raised temperature range, improving free-cooling. In fact the Tennessee climate can allow for more than four thousand hours of water-side economizers (17°C supply water). That leads to a 45% reduction in chiller use.

The new technology (water-side economizers) cannot be retro-fitted to the existing cooling towers.

2.5 Chapter Summary

The following needs and requirements from IT side have been observed:

- Energy requirements are still increasing towards a goal/"acceptable" limit of 20 MW for exascale computers.
- Variability (large swings) of power consumption is becoming more and more important.
- Energy monitoring should be improved using sensors for thorough power measurement and software that will enable pragmatic optimization of consumed power. Energy consumption of all components of the computer system should be monitored.
- Reduce power consumption by choosing, when possible, optimal frequency per application (needs to be done automatically) or force unused nodes to go into deep sleep mode (or even shut them down). Need for energy-aware job management.
- Increasing operation temperature of components saves on cooling but may lead to increase in the power consumption of components (increase of leak currents).
- Good practice to inform the users about consumption from their applications – toward energy allocation instead of cycle allocation?

General considerations:

- Exascale is not far away (2020- 2025) in terms of lifetime of buildings/infrastructures: new building or major refurbishments need to be done for being able to host exascale systems. This includes the need for expandable power supply. Of course, it is likely that a variety of approaches will be used for addressing the exascale challenge from the infrastructure point of view.
- Holistic approach is needed for improving energy efficiency. This holistic approach should, in some cases, go beyond the data centre (best practices for district heating, to think globally about the usage of the water). Building measurement and automation is very important.
- Establish a closer connection between facility people and system staff in order to deal with the swings in load that the systems produce. Monitoring power usage better from a plant side to get an accurate picture of what systems are drawing at various levels. These two parties need to work more closely together in order to derive baselines for system idle/full load and share information. A lot of systems have temperature/flow/pressure sensors that could be of interest to facility people.
- Lacking competence in infrastructure for HPC centers –> training/internships required or a good practice.
- Preparation is very important.
- Floor space for infrastructure equipment = 2 * floor space for IT equipment.
- More than 2000 kg/m2 floor load bearing

Cooling:

- Free-cooling is a general trend. Should be decided after careful analysis of local temperature – free cooling is often possible most of the time, but chillers may be needed for a few days/weeks.
- Direct liquid cooling is used in more and more sites. Good for temperature of components. Be careful about water quality. Be careful about remaining air-cooled components (may suffer from high temperature).
- Air cooling is still attractive even for big sites. Enclosed corridors are needed.

- Heat-reuse is easier with direct liquid cooling (higher temperature) but also possible with air cooling.
- Consider reduction of water usage as a global goal. It may happen that the water consumption needed to produce the electricity for chillers is larger than the one used for evaporation systems.
- Large swings in return temperature may lead to the need of specialized chillers, or an ability to mitigate large swings through additional mixing, baffles, or containers.

Power-supply:

- UPS only for critical equipment is the rule
- Use higher voltage for distribution (US=480VAC, with interest in 600VAC). Reduce length of power cables between the transformers and the racks in order to reduce electricity losses.

Indicators:

- PUE is widely used but is not enough since it doesn't take into account heat recovery and that part of the infrastructure may be part of the IT equipment (rear cooling doors).

3 Energy efficiency in HPC

Keeping high-performance computing (HPC) affordable and cost effective has always been a key requirement. Increasing energy efficiency in HPC systems will reduce energy consumption, decrease generation of heat, lower operational costs, and improve system reliability. Energy conservation and minimized operating costs have become increasingly important factors shaping how HPC resources are used today. Key market players are working on management systems and techniques for this purpose.

The growing energy demands of high performance computing systems require new approaches across the energy ecosystem. From data center design, to system cooling and power distribution, all the way down to the individual HPC applications - researchers and engineers are bringing a wealth of new ideas and technologies to help to address these challenges.

Hereafter, we present first the main findings from ISC 2013 in Leipzig and the HPC workshop in Lugano, in relation to monitoring and administration systems that contribute to energy efficiency in HPC. Then, we present new information collected directly during meetings and discussions with vendors regarding energy efficiency, including for cooling and for electricity supply.

3.1 State of the art in monitoring and administration systems

3.1.1 IBM

The IBM Systems Director Active Energy Manager is the cornerstone of IBMs energy management framework. It measures, manages, monitors power and thermal energy usage and also integrates with infrastructure and enterprise management suites.

As supercomputers scale to millions of cores to reach the Exaflop/s performance, the underlying resource management software architecture needs to provide a flexible mechanism for a wide variety of workloads executing on the supercomputer. IBM's LoadLeveler with Energy aware scheduling addresses this requirement by setting optimal processor frequency on the set of nodes where a job runs or setting the node frequency at lowest power consumption when no job is scheduled.

LoadLeveler provides the capability to develop energy saving functionalities. Using this energy functionality, a job can run with a lower CPU frequency to save energy or run faster than default typically at the expense of using more energy.

A user or administrator can set the acceptable performance degradation or required energy saving in a job. LoadLeveler will choose a suitable CPU frequency for the job. The use of energy policy tag helps LoadLeveler identify the energy data associated with a job. With the energy data, LoadLeveler can decide which frequency should be used to run the job with minimal performance degradation. The energy data includes:

- Power consumption and the elapsed time when run with default frequency
- The estimated power and energy
- The elapsed time at other frequencies
- The percentage of performance degradation (w.r.t. runtime)

Setting the energy policy tag in the job command file, the energy data will be generated and stored in the database when running the job for the first time. If the job is submitted again with the same energy policy tag, the same policy will be used. Submitting jobs using a new

energy function for the first time, it must be taken care of keeping the tag name unique among the tags previously generated.

IBM, L. Brochard, IBM STG - Deep Computing

Enabling exascale computing seems to be relying on three main pillars: effective computing units, effective infrastructure and effective management. At the moment there is a widespread assumption that the future exascale system must operate within 20MW power limit. None of currently used technologies allows for getting even close to the desired values. There are however different areas where we can see movement towards the goals. These are: new computing units that are featured by increasingly higher Gflop/s/Watt ratios. Cell, GPUs, MIC are all examples of solutions developed over time that are characterized by increased computational power while keeping the power at the same level. Data management is a fairly new area that is not yet as developed as the computing units, however there is some development in this area as well (3D memory, SSD etc.). Right cooling technologies are one of the key aspects of energy efficient computing. IBM presented comparison between traditional, air cooled solutions, rear-door air cooled one and solution based on direct liquid cooling (IBM iDataPlex dx360 M4). The liquidcooled version allows for energy savings both on the cooling infrastructure and servers, due to better environmental conditions. Direct liquid cooling proved to be the most efficient cooling method allowing for increasing the inlet temperature of the coolant to 50°C and thus enabling all-year round free cooling for such machine. There is however a drawback, starting from certain higher temperature points the power consumed by the chip increases significantly.

IBM Platform HPC

IBM Platform HPC is a complete high performance computing (HPC) management solution in a single product. It includes a rich set of out-of the-box features that empowers high performance technical computing users by reducing the complexity of their HPC environment and improving their time-to-solution.

IBM Platform HPC allows technical computing users in industries such as manufacturing, oil and gas, life sciences, and higher education and research to deploy, manage and use their HPC cluster through an easy to use web-based interface, thus minimizes the time for setting up and managing the cluster for end users and allows them to focus on developing applications rather than on managing infrastructure. Platform HPC provides full cluster management capabilities - from cluster provisioning and management to workload management and monitoring. All the functions required to operate and use a cluster are installed at once and are tightly integrated. The product is designed to deliver faster time to system readiness, ease-of-use and improved application throughput.

Figure 1 IBM Platform HPC architecture

Key capabilities include:

- Comprehensive, easy-to-use cluster management
- Next generation, easy to use interface
- Integrated application support
- User-friendly, topology aware workload management
- Robust workload and system monitoring and reporting
- Dynamic operating system multi-boot
- GPU scheduling, management and monitoring
- Robust commercial MPI library

With Platform HPC, the master node is aware of resource demands of the jobs queued and what each node is doing. It will power down those nodes not being used, aside from keeping about 5-10 percent available at all times. When usage increases, the master node powers on what's needed within a couple minutes.

3.1.2 SGI

SGI Management Center (SGI MC) software provides a comprehensive operational management platform for technical computing. Key features in SGI MC include: robust image management with version control supporting heterogeneous Linux environments; profoundly fast network bootstrap and provisioning; system wide instrumentation, event and alert monitoring with health tracking; component failure-analysis and failure-tracking; power status and utilization monitoring; and power policy management with Intel DCM. Using SGI MC simplifies many administrative tasks with a single management console and accelerates results for the data center. [1]

SGI collaborates with Altair to support energy efficient computing, with a goal to deliver intelligent scheduling based on power consumption. Energy efficiency is being provided by SGI Management Center, which provides a powerful yet flexible interface through which to initiate management actions and monitor essential system metrics for all SGI systems.

Power Option is an optional feature that further extends system management capabilities and adds the capability to dynamically manage electrical load from individual nodes to the entire system from a single point on the system console. Policy driven dynamic power management allows customers to optimize power per computer cycle and operate effectively within facility power constraints and changing electrical rate structures. Fine grained power management allows processes to continue running and improves reliability and response to changes. This option requires underlying hardware support, which is available in Intel® Xeon® 5600-based servers having TY6 or TY15 motherboards, or for most Intel Xeon E5-2600-based servers, including SGI® ICE™ X. SGI and Intel® worked together to achieve this feature. [2]

SGI presented at the HPC workshp in Lugano (April 2013) its ideas for energy-efficient systems on the example of the SGI ICE-X system.

This system features both water-cooling for most power dense parts (CPUs, GPUs or MIC units) and air-cooling for other parts. The machine is delivered in 4 rack islands where each island is self-contained with regards to cooling, providing closed loop for air. The system may be cooled with water up to 32°C providing 40°C outlet water, which is warm enough for direct re-using to e.g. heat up the building. The PUE of the machine isle (288 nodes), without facility infrastructure, is 1.09 for mixed water and air-cooling system.

3.1.3 BULL

Bullx Supercomputer Suite facilitates proficient rule-based power management and provides an effective and flexible way to optimize the overall energy footprint, while avoiding excessive power consumption.

Bullx supercomputer suite is developed upon SLURM Resource and Job Management System, which allows energy consumption recording and accounting per node and job along with parameters for job energy control features based on static frequency scaling of the CPUs. The Resource and Job Management System (RJMS) is the HPC middleware responsible for distributing computing resources to user applications. Appearance of hardware sensors along with their support on the kernel/software side can be taken into account by the RJMS in order to enhance the monitoring and control of the executions with energy considerations. This essentially enables the applications' execution statistics for online energy profiling and gives the possibility to users to control the tradeoffs between energy consumption and performance[3].

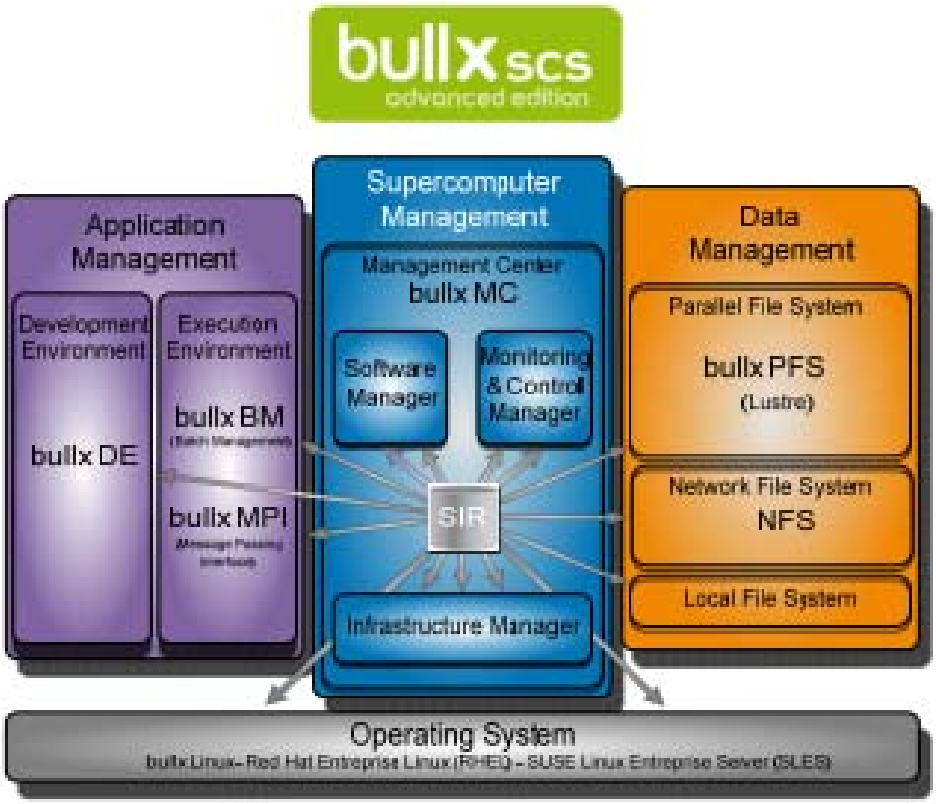

Figure 2 Bullx architecture

Bullx supercomputer suite has a Power Manager [4] which:

- Provides a tool to implement power policies
- Is easy to customize
- Has an event correlation tool: SEC (Simple Event Correlator) which:
  - Parses motor for multiple inputs
  - Is capable of performing complex comparisons
  - Once the pattern is matched, an action is taken
  - Acts upon predefined rules (Example if a certain rack has a very high temperature, PowerManager can tell Bullx BM not to send jobs to nodes on a specific rack)

Bull, L. Cargemel (HPC workshop in Lugano)

Bull's presentation covered last 8 years of HPC evolution, depicting the increasing energy and computation density of the HPC systems. Starting from 10kW/rack and 100TF in 2005 Bulls vision ends in 2015 with a 120PF system with a total power budget of 10MW. The power density of this future machine is predicted to be 150 kW/rack of purely liquid cooled machine.

Bull's vision of an exascale machine is an evolution of currently used technologies but refined, made more efficient and deployed at larger scale. The reasonable limits for the Exaflop machine will be 20MW power and 600 square meters of floor space. The most important topics that need to be solved are related to the increasing complexity of the future systems. Bull's solution to increase reliability of the future systems is implementation of distributed checkpoint / restart on fast, non-volatile memory.

The current generation of Bull's liquid cooled machines was presented as an evolutionary step that will lead to future exascale machines. The machine can host up to 288 sockets per rack

and can be cooled with inlet water at 35°C. The next generation should be able to operate with at least 40°C to increase the energy re-usability. The future machines are predicted to be more compact, the compute nodes and interconnect should be integrated into racks that will be treated as self-contained islands. The copper will stay as the dominant medium connecting the elements inside the rack while the rack-islands will be connected with high bandwidth optical connections

Another important prediction presented is the increased weight of the computer racks, which may exceed 2000 kg per rack. While it does not influence directly power efficiency, it does affect the requirements on the physical parameters of the future data centers.

3.1.4 BRIGHT

Bright Cluster Manager has a variety of Comprehensive Monitoring features.

According to [5], Bright Cluster Manager, can collect, monitor, visualize and analyze a comprehensive set of metrics. Examples include CPU, GPU and Xeon Phi temperatures, fan speeds, switches, hard disk SMART information, system load, memory utilization, network metrics, storage metrics, power systems statistics, and workload management metrics. Custom metrics can also easily be defined. Metric sampling is done very efficiently — in one process, or out-of-band where possible. There is full flexibility over how and when metrics are sampled, and historic data can be consolidated over time to save disk space.

Bright Cluster Manager (ver6.0) is used to reduce operating costs with its new option to power down unused nodes and system components. This feature is available with either PBS Professional or SLURM [6]. Bright Cluster Manager works closely with PBS Professional to enable automatic shutdown of unused system resources: nodes, disk drives, switches, etc. Generally speaking, every watt saved from the cluster saves an additional watt in terms of cooling.

Additionally, Bright Cluster Manager has power management features [7], which are:

- Power Distribution Unit (PDU) based power control
- IPMI-based power control (for node devices only)
- Custom power control
- HP iLO-based power control (for node devices only).

Power operations, in Bright Cluster Manager, may be done on devices from either cmgui or cmsh. There are four main power operations that can take place:

- Power On: power on a device
- Power Off: power off a device
- Power Reset: power off a device and power it on again after a brief delay
- Power Status: check power status of a device.

The monitoring system of Bright Cluster Manager collects also power-related data. Monitoring power consumption is important since electrical power is an important component of the total cost of ownership for a cluster.

The Bright Cluster Manager, which is mainly tested on APC units, can also provide status information on non-APC power units using custom power scripts. [8].

3.1.5 HP

HP Insight CMU is the HP monitoring system that serves as a powerful tool for installing Linux software images, including middleware such as Message Passing Interface (MPI) and job schedulers. HP Insight CMU can be used in order to manage a number of standalone systems that are similar in hardware and software configuration.

The monitoring feature of HP Insight CMU is designed to synthesize environmental, performance, and administration information from the cluster for the IT manager. It has an optional application interface, which is designed for easy extensibility of the scalable framework. This interface has been used to integrate several popular HPC management products with HP Insight CMU. HP Insight CMU has been integrated with the Altair "PBS Professional" scheduler imaging technology for dynamic provisioning ("Green Provisioning"). Integration has also been done with HP Insight CMU's remote management features so that PBS Professional can monitor, shutdown and restart HPC systems as necessary, to ultimately manage power use to save-reduce energy consumption. [9]

According to Altair website [10] and validating this info by several leading website, Green Provisioning claims that has lowered their energy use by up to 30 percent. Main features and benefits of Green Provisioning are:

- **Dynamic system monitoring:** Monitors the state of the queues and the activity leveli on the system nodes and makes decisions to power down nodes that are running no jobs or if queuing jobs require powered down nodes to be booted up.
- **Customizable power reduction methods:** Power use adjustment can be customized to suit the computing environment. Throttle CPU frequency to minimize idle power consumption, or fully shutdown nodes to eliminate consumption entirely.
- **Node power down priorities:** System Administrators have the ability to prioritize which nodes should be powered down in what order by assigning a "power down priority" to each node. This way one can ensure the highest performance nodes are available as much as possible,
- **Node power cycle time delays:** All system nodes have their own parameters that control how long the node should stay idle before it is powered down and how long to wait for the node to boot up.
- **Temperature-based placement:** Schedule jobs to nodes based on temperature to reduce cooling costs by evenly spreading out the A/C loads.
- **Custom resources are supported:** Fully supports node-level custom resources. When working out which powered down nodes should be booted up, the custom resources requested by each queued job are matched against the resources available on each node.
- **Multi-node jobs are supported:** Fully supports jobs that request multiple nodes or multiple chunks.

3.1.6 CRAY

Cray CS300 - Advanced Cluster Engine (ACE) has a flexible and energy-efficient Architecture. Cray delivers a complete end-to-end solution combining hardware, software and professional services to support the job execution, monitoring, management and debugging tools that facilitate running large, complex HPC applications.

Main specifications of the System Administration, Resource Management and Job Scheduling of ACE are the integrated job scheduling and resource management with options for Grid Engine, SLURM, Altair PBS Professional, IBM Platform LSF or Torque/Maui where some of those offer advanced power based management functionalities as discussed above. Additional capabilities are the fine-grain system power, the temperature monitoring and the remote power control [11].

Cray, P. Williams (HPC workshop in Lugano)

The Cray presentation was focused on discussing open issues that have to be solved before the Exascale becomes reality. The key problems are:

- power. Current approach to powering and securing the power are not good enough
- concurrency. Parallelism is the key that will allow us to reach the exascale.
- programming difficulty. Concurrency makes programming a huge issue, increasing the concurrency makes the problem worse. New programming paradigms must be created to hide the complexity from the programmers
- resiliency. Future machines will be more complex than the ones we are using now.

With the growing complexity grows also failure rate. Measures must be taken to cope with this.

According to Cray MPI will stay as the most important way of programming for large problems.

3.1.7 Energy Efficient HPC Working Group Natalie Bates

The Energy Efficient HPC Working Group [41] gathers entities involved in development of the HPC technologies with special attention to the topics related to energy efficiency. The working group acts as a forum for sharing information, best practices, guidelines and recommendations between partners.

One of the most important topics is power measurement methodology. While there are some indicators used by vendors or organizations to measure the energy efficiency (Top500 power consumption, Green500, PUE etc.) the informational value of these indicators is very limited due to lack of standards and procedures used while measuring the values. The working group proposed a set of recommendations that cover most important topics: how to measure, what to measure and where to measure. There are three tiers of power measurement quality, each one more difficult from a technical point of view but producing higher quality results. The rules and recommendations were included by the green top 500 to the submission procedure.

Apart from the methodology the group gathers also best practices regarding the instrumentation of the data center, the rules of procurement of new servers and data center infrastructure that should result in an infrastructure that is more energy efficient and instrumented well enough to provide tools for gathering comparable data.

The group proposes a set of new metrics which were presented (itUE and TUE), which may substitute non-accurate PUE.

The group is closely following the development of the HPC market and responded to the recent trends by introducing new team that is focusing at commissioning liquid cooled systems.

3.1.8 Ganglia

Ganglia [12] is a scalable distributed monitoring system for high-performance computing systems such as clusters and Grids. It is based on a hierarchical design targeted at federations of clusters. It leverages widely used technologies such as XML for data representation, XDR for compact, portable data transport, and RRDtool for data storage and visualization. It uses carefully engineered data structures and algorithms to achieve very low per-node overheads and high concurrency. The implementation is robust, has been ported to an extensive set of operating systems and processor architectures, and is currently in use on thousands of clusters around the world. It has been used to link clusters across university campuses and around the world and can scale to handle clusters with 2000 nodes.

Ganglia is a BSD-licensed open-source project that grew out of the University of California, Berkeley Millennium Project which was initially funded in large part by the National Partnership for Advanced Computational Infrastructure (NPACI) and National Science Foundation RI Award EIA-9802069. NPACI is funded by the National Science Foundation and strives to advance science by creating a ubiquitous, continuous, and pervasive national computational infrastructure: the Grid. Current support comes from Planet Lab: an open platform for developing, deploying, and accessing planetary-scale services.

It is based on a hierarchical design targeted at federations of clusters. Federation is achieved using a tree of point-to-point connections amongst representative cluster nodes to aggregate the state of multiple clusters. At each node in the tree, a Ganglia Meta Daemon (gmetad) periodically polls a collection of child data sources, parses the collected XML, saves all numeric, volatile metrics to round-robin databases and exports the aggregated XML over a TCP socket to clients. Data sources may be either gmond daemons, representing specific clusters, or other gmetad daemons, representing sets of clusters. Data sources use source IP addresses for access control and can be specified using multiple IP addresses for failover. The latter capability is natural for aggregating data from clusters since each gmond daemon contains the entire state of its cluster.

3.1.9 Pinguin computing - Scyld ClusterWare

Scyld ClusterWare is an HPC cluster management solution compatible with the Linux distributions RedHat Enterprise Linux and CentOS. Scyld ClusterWare is designed to make the deployment and management of a Linux cluster as easy as the deployment and management of a single system. Functionality:

- Super-fast cluster provisioning
- Single System Image Architecture that guarantees configuration consistency
- Support for internal clouds and cloud bursting
- Web-service-based architecture for management and workflow submission from anywhere
- Qualification and optimization for Penguin hardware for an optimal user experience
- Certification as Intel Cluster Ready reference architecture for SSI clusters

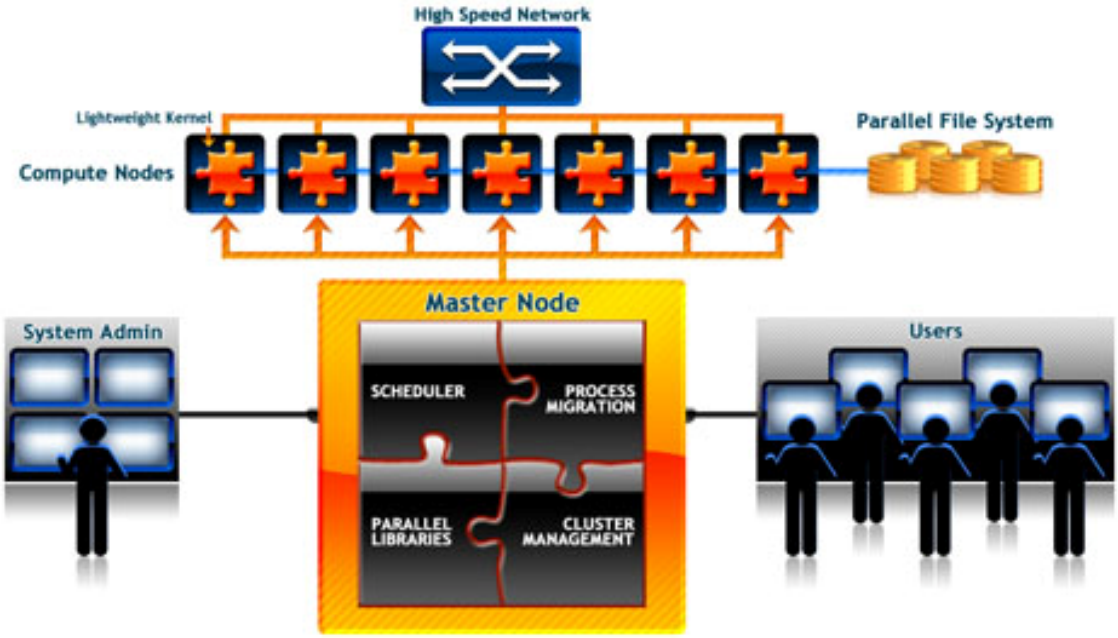

Figure 3 Scyld ClusterWare architecture

Advantages of using Scyld ClusterWare:

- Immediate productivity with a complete, simple to install and integrated cluster management solution.
- Extreme computational needs require an industry tested, robust, scalable and standards compliant computing solution.
- Helps effectively manage this complexity through a "single system" view.
- Consistent cluster configuration is guaranteed through Scyld's 'single system image' .
- User setup, library installations, shared network mounts and resource monitoring only need to be performed on the master node.
- Compute nodes are added to the cluster simply by plugging them into the network and powering them on, simplifying configuration and guaranteeing consistency across the cluster.
- Scyld ClusterWare supports both diskless and local OS installations through the Hybrid functionality, allowing you to create heterogeneous cluster configurations

Scyld Insight is a web-service-based cluster management and monitoring GUI. With Scyld Insight, cluster administrators can monitor system metrics that provide insight into a cluster's health, activity and utilization in real time. They can configure Scyld Clusterware and quickly analyze any set, without needing a high level of HPC cluster expertise.

3.1.10 Nagios

Nagios [13] is a monitoring system that enables organizations to identify and resolve IT infrastructure problems before they affect critical business processes.

Designed with scalability and flexibility in mind, Nagios gives you the peace of mind that comes from knowing your organization's business processes won't be affected by unknown outages.

Nagios is a powerful tool that provides you with instant awareness of your organization's mission-critical IT infrastructure. Nagios allows you to detect and repair problems and mitigate future issues before they affect end-users and customers.

Functionality:

- Plan for infrastructure upgrades before outdated systems cause failures
- Respond to issues at the first sign of a problem
- Automatically fix problems when they are detected
- Coordinate technical team responses
- Ensure your organization's SLAs are being met
- Ensure IT infrastructure outages have a minimal effect on your organization's bottom line
- Monitor your entire infrastructure and business processes.

Nagios provides health and status summary information for all the servers, routers, switches, power devices and related services you are monitoring. It provides alarms with notification mechanisms that alert you of potential problems, highlight critical events that require immediate attention and escalate the alerts if not rectified.

Nagios can be configured for more extensive distributed monitoring where the overhead of performing service checks from a central monitoring server is offloaded onto one or more distributed servers. While small to medium sized shops may not need such an environment, when you start monitoring hundreds or even thousands of hosts this becomes quite important.

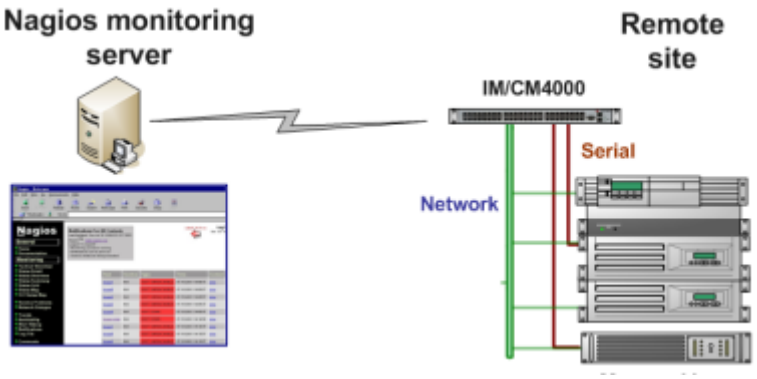

Figure 4. Nagios distributed monitoring

At the remote site, the Opengear console server functions as a distributed Nagios server and performs checks on all the hosts and services that you define for that site. A Nagios NSCA client and NRPE server are embedded in the console server so it can perform these checks. These programs enable scheduled check-ins with the central Nagios monitoring server and send check results across the network to the remote monitoring server.

3.2 Cooling systems and their efficiency

3.2.1 Trends in HPC Cooling

HPC systems consume more and more energy. This implies that more and more emphasis is put on the cooling system. Several trends can be observed. One trend is to use more free cooling. This can be done by using water at some point in the loop or using air-cooling more or less directly to the outside air. Where water is used in the cooling system the trend is to capture the heat as early as possible. Traditionally, Computer Room Air Conditioners (CRAC units) have been used and the air in the room has been a mix of hot and cold air. To get good efficiency of such systems it is essential to have as high temperatures as possible at the cooling coil. To achieve that, hot and cold isles were introduced. This was partially successful but cold and hot air could still mix to some extent. To get even higher efficiency encapsulated systems were introduced where the hot and cold air where separated by some kind of encapsulation. This increased the efficiency even more. Similar ideas were introduced in cooled rack doors where the cooling coil was placed directly at the exhaust of the hot air in the rack. To get even higher efficiency direct liquid cooling is the most recent trend for HPC systems. Thanks to these solution's unique capabilities a higher energy-efficiency of both systems and whole data centers can be achieved. Thanks to the high temperature water generated as a by-product of the cooling process it becomes possible to re-use the waste heat.

3.2.2 Direct Liquid Cooling

Direct Liquid Cooling is an old technique, which recently has been re-introduced on a broad scale by several companies including the major HPC providers. With "Direct Liquid Cooling" we refer to using some liquid (often water) more or less close to the CPU/memory etc. to cool a computer system.

This has several advantages:

- Low extra energy used for cooling in pumps or fans etc. compared with air- cooling.
- Efficient and more precise control of cooling – possibility to run CPU faster.
- Higher outgoing water temperature than going through air.
    - This gives the possibility to use free cooling for longer time of the year and in more locations.
    - It also open more possibilities for heat re-use.
- It makes possible higher energy density.

There are however some issues related to this kind of cooling such as the lack of standards for water temperatures, for water quality, for connectors used and for the increased weight of the rack etc. that may slow down the adaptation of thess kind of solutions in existing data centres.

Lack of standards regarding the environment conditions on which liquid cooling operation may cause issues during the commissioning phase. Different machines are featured by different physical dimensions, different flow requirements, and different operating temperatures. There might be even problems with purity levels of the water used by different vendors. All this differences may cause lowered competitiveness of different vendors as choosing one product may require customized adjustments to the facility infrastructure. Standardizations efforts are however pursued by organizations like ASHRAE.

3.2.2.1 Different Types of Direct Liquid Cooling

Direct liquid cooling can be more or less efficient. First it depends on which components are liquid cooled and which are cooled by ambient air. The CPU is the component with largest power need and the CPU is the first candidate for liquid cooling. However, other components like memory, power supplies may or may not be liquid cooled. In the case that some components are not liquid cooled the computer centre may still need some air-cooling, which makes the efficiency lower. Different cooling components can also be differently optimized to transfer the heat. It seems that cooling components, which are more specialized for a certain component are slightly more efficient than liquid cooling devices, which are more generic. Many companies have produced liquid cooling devices for the enthusiast market ("over clocking") for some time and are now starting to look at the HPC market. The typical HPC user is however expected to wait for the larger companies to embrace direct liquid cooling. This seems also to be the case, at the latest International Supercomputer Conference ISC13 most major HPC companys revealed systems with liquid cooling. IBM already has systems both in Blue Gene line and with Intel CPUs that are directly liquid cooled. Bull showed a flexible system for liquid cooling where the components were attached to a cooling plate. The system is modular and should be easy to adapt to new processors and GPUs. Many smaller companies displayed components for liquid cooling which can be used by the system integrators. Cray has however chosen another approach where they have cooling coils in racks adjacent to the CPU racks. This solution is quite similar to liquid cooled doors discussed above.

3.2.3 Best practices in air cooling

Although recently there has been significant increase in the deployment of a variety of advanced cooling solutions based on water or liquid cooling, the use of air cooling in data centers is still a viable alternative, especially where the deployment of new computational resources is gradual or where the target is not very high power density.

In the following we outline some best practices and recommendations about efficient organization of air cooling that were kindly shared with us by representatives of Fujitsu.

The importance of organization of short, straight air flow paths is underlined by the fact that doubling air flow may increase eight times the power used by fans, thus dramatically decreasing the data center infrastructure efficiency.

That is why a major goal of organizing airflow is to prevent the mixing of cold and hot air and to make sure that the cold air can efficiently reduce the temperature of the hot circuitry. Some of the established best practices for the design of air cooling include: perforated tile placement, blanking panels, sealed walls and ceilings, underfloor fans to direct air flow. The newer solutions that aim to improve cooling include in-row cooling, in-rack cooling and even in-chassis cooling, i. e., focusing the cooling near the equipment, sometimes using liquids. The technique of hot- and cold-aisle containment, concentrated on preventing the mixing of hot and cold air, is widely used.

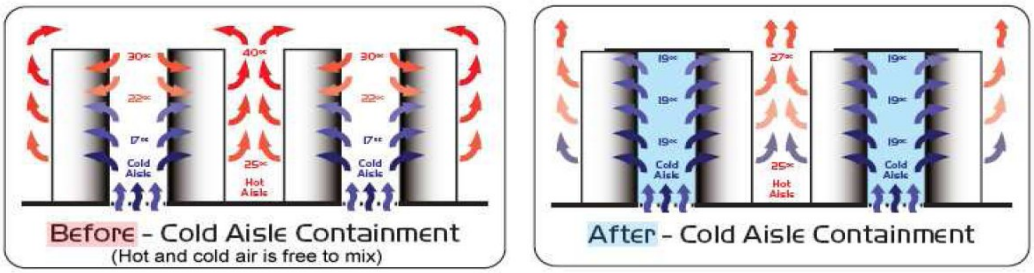

Figure 5 Cold Aisle Containment

The most common solution for a data center that uses computer room air conditioning units and a raised floor calls for the careful arrangement of racks into hot-aisle/cold-aisle configuration where the cold-air inlet side of all equipment racks faces the same direction. If cold-aisle containment is used, the cold-air supply from the cold aisle is held there, so that the only way it can mix with hot air is to go through the equipment racks. **Figure 5** shows the results of applying cold-aisle containment. Such configuration maximizes the efficiency of air cooling, while being relatively easy to deploy in existing data centers. That is why it is recommended for data centers that are not using in-rack or in-chassis cooling. In a similar way the deployment of hot-aisle containment offers benefits in efficiency at a reasonable cost. By directing the exhaust air outside of the data center it can be ensured that resources are not wasted for its cooling inside. For existing data centers both isolation technologies should be evaluated for cost and impact, while for newer data centers they should both be considered if in-rack or in-chassis cooling is not already envisaged.

One important consideration when deploying such isolation solutions should be to make sure they do not prevent the proper operation of the fire detection and fire suppression systems in the data center.

3.3 Electricity

3.3.1 Infrastructure Technologies

Two novel technologies used to improve the energy efficiency of UPS (uninterruptible power supply) in data centres were presented at the third session of the 4th European HPC Centre Infrastructure Workshop. The first technology, presented by Simona Vrabiescu of Maxwell Technologies, is the use of Ultracapacitors (electrostatic energy storage) in substitution of batteries. Lorenzo Giuntini of General Electric, presented two new technologies: IEMi (adapting UPS capacity to load) and eBoost (multi-mode UPS).

3.3.2 Ultracapacitors

Maxwell Technologies is the leading manufacturer of ultracapacitors, a new type of electrochemical capacitors designed to amplify the energy density of traditional capacitors beyond their current limits. These ultracapacitors provide electrostatic energy storage with no moving parts, using non-contaminant materials: carbon, aluminium, electrolyte and paper. They greatly outlive traditional batteries with between 100k and 1M charge/discharge cycles, which provides for around 5-15 years of use. They also achieve much higher power densities (up to 7.5 kW per 1U rack unit), allowing for very fast charging and discharging.

Products based on ultracapacitors include standard cells and multi-cell modules for larger capacities, for use in applications ranging from electronics to renewable energy plants. Initial cost of ultracapacitors is much higher than other UPS technologies (batteries and flywheels), but their long life and low maintenance (only simple voltage monitoring) is supposed to bring total cost of ownership below that of its competitors. The multi-cell modules can be connected in series (up to 750 V) or parallel.

3.3.2.1 Practical use of Ultra-capacitors at CEA

Energy efficiency is a strong concern for the two facilities, because of the large increase of power consumption of world-class supercomputers in the last 10 years, which nowadays typically reaches several MW.

In this context, part of a global approach for improving the energy efficiency (for which CEA received an award of the European Code of Conduct for Data Centre in May 2013), CEA is implementing best practices aiming at reducing the losses in the electrical distribution as documented in a recent PRACE white paper [14].

This includes:

- limiting the usage of UPS to critical parts of the IT equipment (mainly storage and network) in order to limit the electrical losses in UPS (the efficiency of a UPS is typically around 90% when used in an HPC centre)
- using ultra-capacitor modules (UCM) for all other components in order to deal with short-term interruptions (less than 150 ms) which occur several times a year.

This UCM solution was first implemented with Tera-100 which, in June 2011, was number 6 of the Top500 list and again with Curie which, in June 2012, was number 9 of the Top500. For reference, a similar solution is used for the Helios supercomputer (in Japan) (number 12 of the Top500 in June 2012).

The technology used for these three systems was developed by Bull; ultra-capacitors are integrated in the power supplies of the compute nodes and are able to provide power during up to 300 ms current outages at full load. The energy efficiency of ultra-capacitors is around 99.8% which means that, for IT equipment using 1 MW of electricity, it is possible to save up to 100 kW of power losses in UPSs.

The experience of using ultra-capacitors at CEA is based on two generations of UCM:

- UCM for fat nodes with a peak power consumption of 1 kW (Tera-100)
- UCM for enclosures with 18 thin nodes, with a peak power consumption of 7 kW (Curie).

This experience shows that UCM are very reliable and effective. The observed autonomy is more than 300 ms since this figure is based on the node power consumption at full load. Since the start of operations of Tera-100, all short-term power interruptions (5 to 6 a year, less than 150 ms) have been dealt with by the UCM without any problem.

For the next generations of UCM, the capacity is expected to reach 12 kF in order to deal with the increasing power usage of nodes. In addition, in order to increase the autonomy of the UCM, automatic reduction of the frequency of the processors in case of short-term power failure is planned.

In conclusion, CEA is fully satisfied with the mix of UPS (for 20% of the load) and ultra-capacitors (for 80% of the load) in terms of operation and energy efficiency.

3.3.3 IEMi and eBoost

The products presented by General Electric do not revolutionize the world of UPS, but instead provide an evolutionary improvement in efficiency by tuning current technologies. Intelligent Energy Management integrated (IEMi) provides load-balancing to parallel-redundant UPS installations to improve efficiency during low system load. Through their Adaptive Capacity Control technology, IEMi maintains the reliability of a double-conversion redundant configuration while moving UPS operating conditions to optimal levels when load is low. The load-balancing can be configured based on calendar and external input from monitoring.

The second General Electric product, eBoost, is a high-efficiency alternative to double-conversion in traditional UPS products. With eBoost, the UPS may optimize efficiency by autonomously selecting the operating mode depending on the quality of the input utility (Multi-Mode UPS) , while output voltage is maintained in compliance with the most common

PQ standards and the requirements of IT equipment PSU. This means the UPS can switch extremely quickly (< 3 ms) between normal and bypass mode, which limits PSU inrush current.

3.4 Other related trends

3.4.1 Intel keynote speech at ISC 2013

Intel is aware of the challenges standing between them and an exascale machine. This includes energy efficiency, packaging density, and new memory models. Intel has plans to address all of these.

The energy efficiency is already a big thing for Intel as the mass consumer market shifts from PC to mobile devices. The last three generations of Intel processors proved to show increased performance while reducing consumed power. Intel's approach seems to be a strategy of small steps: no major revolutions are predicted. Current generations of products (Xeon, Phi) will be continued in traditional tick-tock cycle. The trend for the future will be continuation of huge Xeon cores that will be able to run scalar code while highly parallel load should be executed on Phi cores. As a consequence future HPC products will cover Phi accelerators in their current form; CPU-socket mounted Phi chips for highly parallel codes and hybrids that will feature a mix of X86 cores and Phi-like cores.

Further development steps of the chips will require new materials, new architectures, and a new interconnect. Some examples were presented but all of them are equally uncertain as next generation solutions.

3.4.2 Impressions and trends about processors at ISC 2013

The GPUs are currently considered a normal part of the HPC infrastructure. However, the shortcomings of the GPU-based solutions are becoming obvious and there are several ways of solving these. Because it turns out that having high speed CPUs and GPUs in one cluster proves to be difficult to be used efficiently, one can see a diversification of hardware solutions on the application capability basis.

Writing applications that can run entirely on the GPUs while reducing the CPU role to an application initiator is one example. Usage of alternative, low power chips (e.g. ARM) for computing is considered. While the architecture of the processor does not provide advantage over traditional x86s, low per-unit power consumption and the possibility of using embedded GPGPU units make this solution interesting. There are emerging projects or products focused on this kind of solution (e.g. MountBlanc, HP-Moon Shot, or APU-based SeaMicro servers) that may be an interesting alternative for certain classes of applications.

The era of x86-one-size-fits-all clusters seems to be ending. In order to reach a good efficiency, application capabilities have to be taken into consideration.

3.4.3 Application-Aware Energy Efficiency HPC via Dynamic Voltage-Frequency Scaling (DVFS) at ISC 2013

The energy cost of running HPC systems is growing to a point where it can easily exceed the cost of the original hardware purchase within a few years of operation. This has driven the community to understand how profiles of system's energy usage changes in different types of application workloads and optimize energy costs whenever possible. One way to do this is by

developing an automated framework that uses power and performance models to perform application-aware energy optimizations during execution.

Dynamic Voltage-Frequency Scaling (DVFS) utilization allows CPU speed (clock frequency) reduction and reduced power consumption. In general, different computations have different power requirements and therefore, for computation where the computing element waits for requires the frequency can be reduced to lower power with minimum performance impact. It needs to engage in detailed analysis of a given large-scale HPC application to determine the energy-optimal DVFS settings for each of its computational phases.

Additionally, such solutions may support making CPU clock frequency changes in response to both intra-node and inter-node analysis of the application behavior. For instance, the intra-node approach reduces CPU clock frequencies and therefore power consumption while CPUs lack computational work due to inefficient data movement. On inter-node level the approach reduces clock frequencies for MPI ranks that lack computational work.

In practice, such techniques on a set of large scientific applications are investigated at the San Diego Supercomputer Center on 1024 cores of Gordon, an Intel Sandy Bridge based system. The optimal intra-node technique showed an average measured energy savings of 10.6% and a maximum of 21.0% over regular application runs. Additionally, the optimal inter-node technique showed an average of 17.4% and a maximum of 31.7% energy savings.

3.5 Chapter Summary

Providing highly energy efficient systems is not anymore possible without monitoring applications and management systems. Increasing energy efficiency in HPC systems will reduce energy consumption, heat, lower operational costs, and improve system reliability. Energy conservation and minimizing operating costs have become increasingly important factors shaping how HPC resources are used today.

There are many solutions delivered by hardware vendors, e.g. CRAY (Advanced Cluster Engine), HP (Insight CMU), IBM (**Systems Director Active Energy Manager,** IBM Platform HPC), BULL (**Bullx Supercomputer Suite**) or SGI (**SGI Management Center**).

However, for the heterogeneous environment present in a data center it is worth taking into account independent software solutions like Bright Cluster Manager or open source solutions: Ganglia or Nagios.

Novel technologies are used in electricity and powering HPC infrastructure. They improve the energy efficiency of UPS (uninterruptible power supply) in data centres.

The presented solutions are following:

- Ultracapacitors (electrostatic energy storage) in substitution of batteries, used e.g. in CEA (France),
- IEMi (adapting UPS capacity to load)
- eBoost (multi-mode UPS).

If one of the focuses is to decrease power consumption, ultra capacitors are good candidates to support uninterruptible power.

In order to get even higher efficiency, direct liquid cooling is the most recent trend for HPC systems. Thanks to the high temperature water generated as a by-product of the cooling process, it becomes possible to re-use the waste heat.

There are several advantages of hot hot water and direct liquid cooling:

- Low extra energy used for cooling in pumps or fans etc. compared with air- cooling.

- Efficient and more precise control of cooling – possibility to run CPU faster.
- Higher outgoing water temperature than going through air.

This gives the possibility to use free cooling for a longer time of the year and in more locations.

However, there is still worth to mention, especially for commercial data centers, that air cooled systems are still the most frequent solutions. The main reason is that the energy density of such data centres is lower.

4 Assessment of petascale systems

Considering that the first petascale supercomputer was presented in 2008 and that the first exascale system is projected for 2018, this year 2013 marks the midway point in the "petascale era". Roadrunner, the first HPC system to achieve a peak performance of 1PFlop/s, has now been decommissioned, but there are now 40 supercomputers around the globe that have reached 1 PFlop/s using all sorts of different architectures, vendors, components, and infrastructure.

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

The chapter is divided into two sections:

- **Market Watch 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.

4.1 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 40 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 June 2013 edition of the Top500 List.
- A **static analysis** of the characteristics of the supercomputers contained in the snapshot: architecture, components, performance, and infrastructure requirements.
- A **dynamic analysis** of the evolution and trends in the petascale market based on previous analyses.

4.1.1 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.

We can identify four types of 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.

4.1.1.1 HPC related electronic publications and web sites

- http://www.Top500.org/ - The Top500 supercomputer sites publishes the top 500 list of general purpose systems that are in common use for high-end applications. The present Top500 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 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 [15]. 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 [16] or GARTNER [17] 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 [18], or regular market updates with some predictions and forecast).

4.1.1.2 Computing centre websites

The list of computing centre web sites, obtained from the November 2012 Top500 list, is presented in the Table 1

System	Site	Web Address
Tianhe-2	National University of Defense Technology China	http://www.nudt.edu.cn/index_eng.htm
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
Stampede	Texas Advanced Computing Center - University of Texas United States	http://www.tacc.utexas.edu/stampede
JUQUEEN	Forschungszentrum Juelich (FZJ) Germany	http://www.fz-juelich.de/ias/jsc/EN/Expertise/Supercomputers/JUQUEEN/JUQUEEN_node.html
Vulcan	DOE/NNSA/LLNL United States	https://computing.llnl.gov/?set=resources&page=OCF_resources#vulcan
SuperMUC	Leibniz Rechenzentrum (LRZ) Germany	http://www.lrz.de/services/compute/supermuc/
Tianhe-1A	National Supercomputing Center in Tianjin China	http://www.nscc-tj.gov.cn/en/
Fermi	CINECA Italy	http://www.hpc.cineca.it/content/fermi-reference-guide
Spirit	Air Force Research Laboratory United States	http://www.afrl.hpc.mil/hardware/#spirit
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
Blue Joule	Science and Technology Facilities	http://www.stfc.ac.uk/hartree/42937.aspx

System	Site	Web Address
	Council - Daresbury Laboratory United Kingdom	
Pleiades	NASA/Ames Research Center/NAS United States	http://www.nas.nasa.gov/hecc/resources/pleiades.html
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
DiRAC	University of Edinburgh United Kingdom	http://www.epcc.ed.ac.uk/facilities/dirac
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/
MareNostrum	Barcelona Supercomputing Center Spain	http://www.bsc.es/marenostrum-support-services/mn3
Kraken XT5	National Institute for Computational Sciences/University of Tennessee United States	http://www.nics.tennessee.edu/computing-resources/kraken
Lomonosov	Moscow State University - Research Computing Center Russia	http://parallel.ru/cluster/lomonosov.html
HERMIT	HWW/Universitaet Stuttgart Germany	http://www.hlrs.de/systems/platforms/cray-xe6-hermit/

Table 1: HPC computing centre URLs

4.1.1.3 Vendor web sites

There are a large number of companies that design and produce HPC related hardware and software. The following list of vendors is based on the vendors that supplied the most powerful 50 systems of the June 2013 Top500 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

4.1.1.4 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

4.1.1.5 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.

- **Feed 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 [19] aggregator has been maintained and used, and can be accessed freely. This Netvibes page 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.

4.1.1.6 ***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 engines 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.*

4.1.2 Snapshot

On June 17th 2013, the 41st Top500 list of the world's most powerful supercomputers was presented at ISC13 in Leipzig, and it included 40 systems with a peak performance of at least 1PFlop/s. These systems are described briefly in Table 3 and will be used in the subsequent comparison and analysis of the following section. The list includes all systems which provided the necessary tests of the Linpack benchmark. There is also a very hot discussion worldwide whether Linpack is the best benchmark reflecting the performance of HPC systems (criticism: not possible to show it with only one benchmark). Several other proposed benchmarks aiming to be more appropriate while presenting the performance, reliability, scalability, e.g. Graph500, Green Graph500.

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, 15 systems)
- Lightweight: use many low-power RISC processors (in green, 10 systems)
- Traditional: use only standard high-performance processors (in blue, 15 systems)

System	Site (Country)	Model (processor / accelerator)	LINPACK / peak (PFlop/s)
Tianhe-2	NSCC-GZ (China)	NUDT TH-IVB (Intel Xeon / Intel Xeon Phi)	33.86 / **54.90**
Titan	ORNL (USA)	Cray XK7 (AMD Opteron / NVIDIA Tesla)	17.59 / **27.11**
Sequoia	LLNL (USA)	IBM Blue Gene/Q (IBM PowerPC)	17.17 / **20.13**
K Computer	RIKEN (Japan)	Fujitsu Cluster (Fujitsu SPARC64)	10.51 / **11.28**
Mira	ANL (USA)	IBM Blue Gene/Q (IBM PowerPC)	8.59 / **10.07**
Stampede	TACC (USA)	Dell PowerEdge (Intel Xeon / Intel Xeon Phi)	5.17 / **8.52**
JUQUEEN	FZJ (Germany)	IBM Blue Gene/Q (IBM PowerPC)	5.01 / **5.87**
Vulcan	LLNL (USA)	IBM Blue Gene/Q (IBM PowerPC)	4.29 / **5.03**
SuperMUC	LRZ (Germany)	IBM iDataPlex (Intel Xeon)	2.90 / **3.19**
Tianhe-1A	NSCT (China)	NUDT YH MPP (Intel Xeon / NVIDIA Tesla)	2.57 / **4.70**
Pangea	CSTJF (France)	SGI ICE X (Intel Xeon)	2.10 / **2.30**
Fermi	CINECA (Italy)	IBM Blue Gene/Q (IBM PowerPC)	1.79 / **2.10**
DARPA TS	IBM DE (USA)	IBM Power 775 (IBM POWER7)	1.52 / **1.94**
Spirit	AFRL (USA)	SGI ICE X (Intel Xeon)	1.42 / **1.53**
Curie TN	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**

System	Site (Country)	Model (processor / accelerator)	LINPACK / peak (PFlop/s)
Blue Joule	STFC (UK)	IBM Blue Gene/Q (IBM PowerPC)	1.25 / **1.47**
Pleiades	NAS (USA)	SGI Altix ICE (Intel Xeon)	1.24 / **1.73**
Helios	IFERC (Japan)	Bull B510 (Intel Xeon)	1.24 / **1.52**
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**
DiRAC	EPCC (UK)	IBM Blue Gene/Q (IBM PowerPC)	1.07 / **1.26**
Hopper	NERSC (USA)	Cray XE6 (AMD Opteron)	1.05 / **1.29**
Tera-100	CEA (France)	Bull S6010/S6030 (Intel Xeon)	1.05 / **1.25**
Oakleaf-FX	SCD (Japan)	Fujitsu PRIMEHPC (Fujitsu SPARC64)	1.04 / **1.14**
Raijin	NCI (Australia)	Fujitsu PRIMERGY (Intel Xeon)	0.98 / **1.11**
Conte	Purdue (USA)	HP ProLiant (Intel Xeon / Intel Xeon Phi)	0.96 / **1.34**
MareNostrum	BSC (Spain)	IBM iDataPlex (Intel Xeon)	0.93 / **1.02**
Kraken XT5	NICS (USA)	Cray XT5-HE (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 BL	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**
Big Red II	IU (USA)	Cray XK7 (AMD Opteron / NVIDIA Tesla)	0.60 / **1.00**
SANAM	KAUST (Saudi Arabia)	Adtech custom (Intel Xeon / AMD FirePro)	0.53 / **1.10**
Mole-8.5	IPE (China)	Tyan FT72-B7015 (Intel Xeon / NVIDIA Tesla)	0.50 / **1.01**
Anonymous	Unknown (USA)	HP ProLiant (Intel Xeon / NVIDIA Tesla)	0.46 / **1.14**
Anonymous	Unknown (USA)	HP ProLiant (Intel Xeon / NVIDIA Tesla)	0.42 / **1.08**
Anonymous	Unknown (USA)	HP ProLiant (Intel Xeon / NVIDIA Tesla)	0.29 / **1.05**

Table 3: Snapshot of current petascale systems

4.1.3 Static Analysis

To gain more insight about the specific techniques used to achieve 1 Pflop/s performance, a statistical and graphical analysis of the components, features, and infrastructure of each of the systems is performed. By analysing each characteristic independently and then merge the resulting conclusions, the market can be described in much more detail, providing a better understanding of the underlying environment.

4.1.3.1 Year of construction

Peak performance of 1 PFlop/s was reached for the first time in 2008 (in general purpose publicly listed computers), yet the oldest petascale systems in production today are all from 2010. In fact, more than half of the systems in the market watch (60%) were built or updated in the 2012-2013 timeframe. Obviously, it is still early to conclude anything about 2013, since there are usually more large-scale announcements at SC in autumn, which should bring the total for the year at least as high as 2012 if not higher.

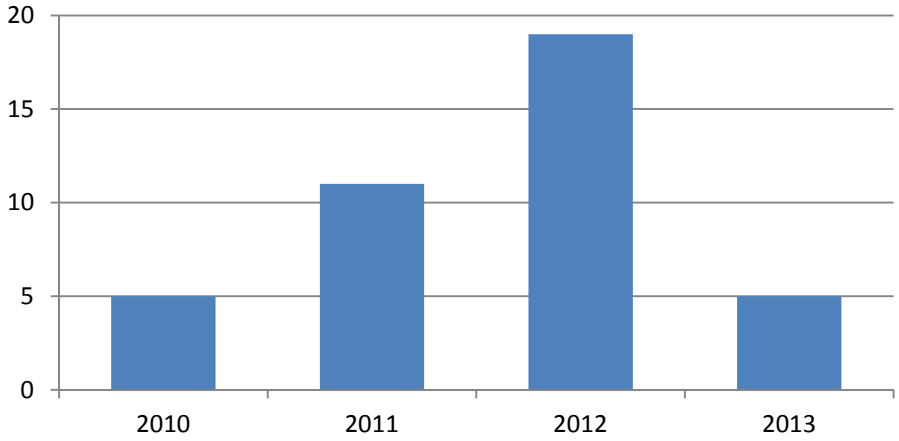

Figure 6: Petascale systems by year of deployment

4.1.3.2 Country

China has once again taken the number one spot on the Top500 List, but the USA still has significantly more petascale sytems (17, almost three times of China's 6) and therefore dominates the market (43%). Japan is third with a 10% share, although only slightly ahead of Germany and France (both 8%). The UK is present with two systems while Italy, Spain, Russia, Australia, and Saudi Arabia each have one. The EU as a whole controls exactly one fourth of all petascale systems, which would be considered second place between the USA and China.

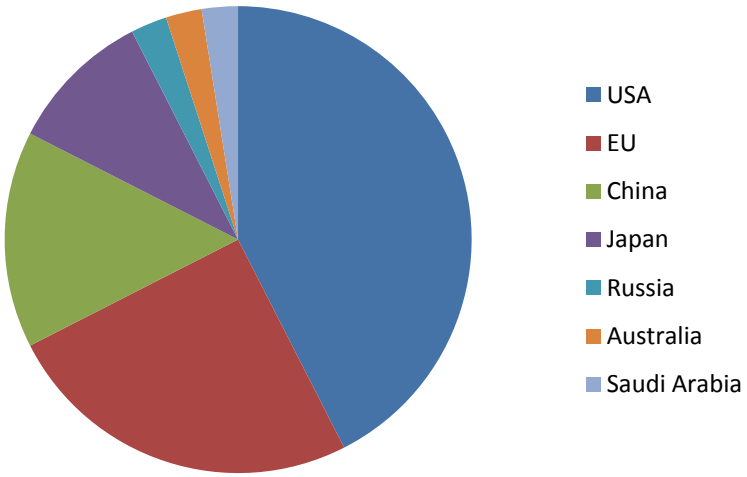

Figure 7: Petascale systems by country

Architecture-wise it is interesting to see that more than 83% of China's petascale computers are accelerated, compared with 41% for American and 25% for Japanese systems. In contrast, none of the members of the EU use accelerators to achieve their petascale performance, preferring traditional clusters (60%) and lightweight MPP systems (40%).

4.1.3.3 Peak performance

As per the definition of this list, all these systems have a peak performance of at least 1 PFlop/s. The maximum theoretical performance is achieved by Tianhe-2 with almost 55 Pflop/s, while the closest to the cut off is Big Red II at exactly 1 Pflop/s. The average for all systems is 5 Pflop/s, yet the median lies at only 1.5 Pflop/s, which means that although half of the systems are in the relatively low range between 1 and 1.5 Pflop/s, the high performers pull the average up (the top 5 systems are at least an order of magnitude above the minimum).

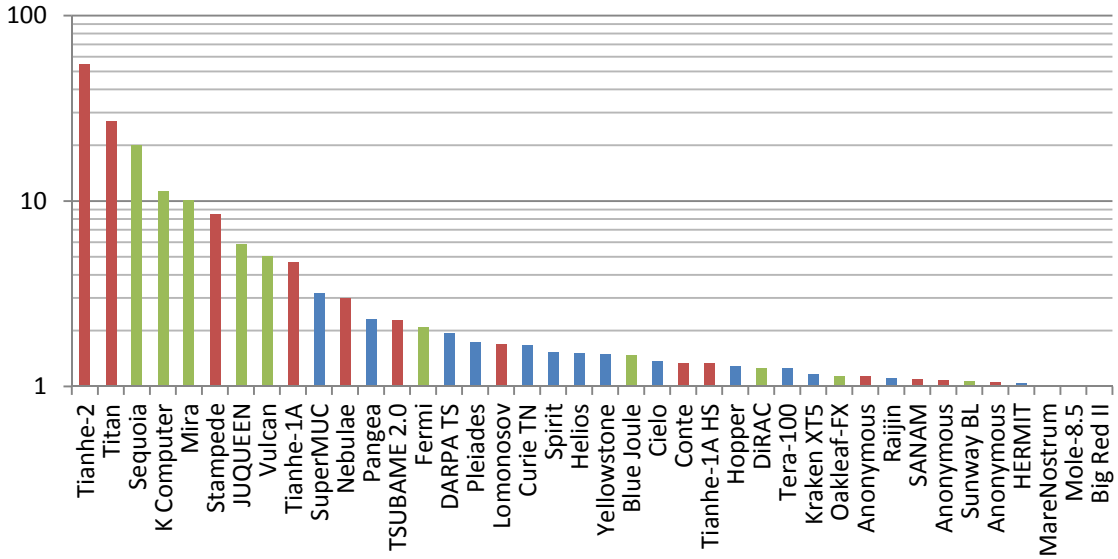

Figure 8: Peak performance of petascale systems (in PFlop/s)

4.1.3.4 LINPACK performance

Real-world performance as measured by the LINPACK benchmark, which is used for ranking on the Top500 List, is obviously always lower than theoretical peak performance, but the difference varies greatly from one system to another. The minimum LINPACK score is just 0.29 PFlop/s, less than one third of the minimum peak performance, while the maximum reaches a little less than two thirds of the highest peak value: 33.86 PFlop/s. As with peak performance, the spread is quite wide owing to the big differences in performance between the top machines and the rest (average LINPACK performance is 3.46 PFlop/s but the median is only 1.21 PFlop/s).

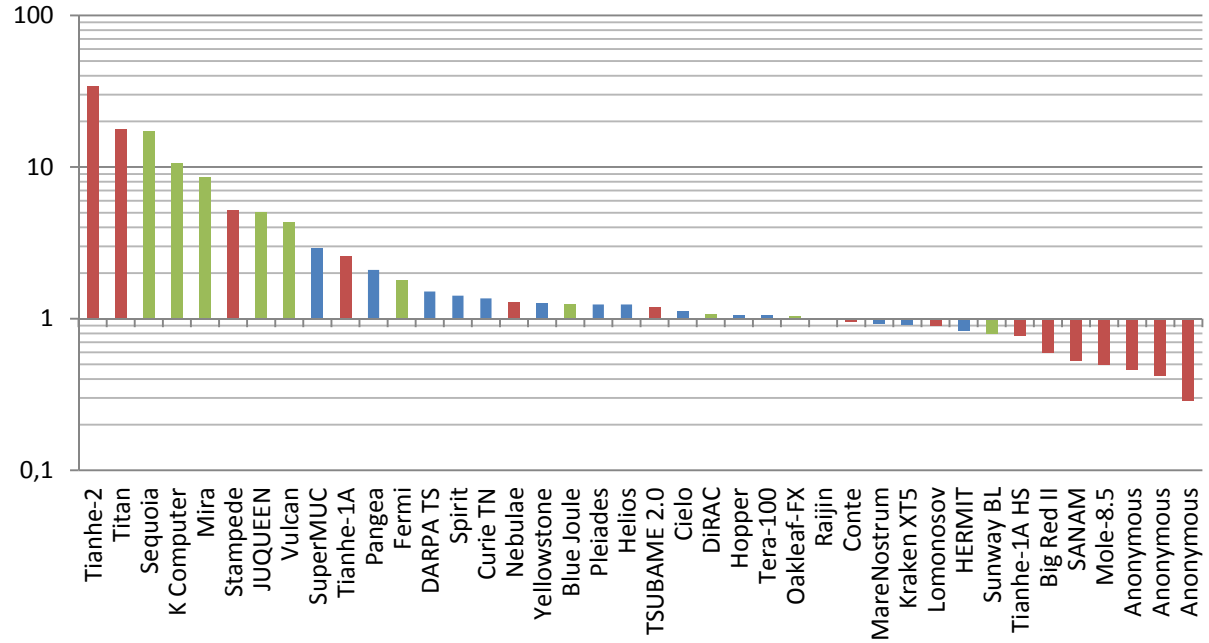

Figure 9: LINPACK performance of petascale systems (in PFlop/s)

4.1.3.5 Vendor

Slightly more than a quarter of all the petascale systems (11 of 40) were built by IBM, while the next most represented vendor, Cray, only manages around half that amount (6 systems that provide 15% market share). HP has 4 full systems (10% share) and one joint venture with NEC (TSUBAME 2.0). Bull, Fujitsu, and SGI each have 3 systems on the list, the same amount as NUDT (National University of Defence Technology), which is a Chinese institution not a commercial vendor (although they partner with Inspur, a Chinese IT firm). The remaining six petascale systems are all from different vendors: Dell, Dawning, T-Platforms, Tyan/IPE, NRCPCET, and Adtech.

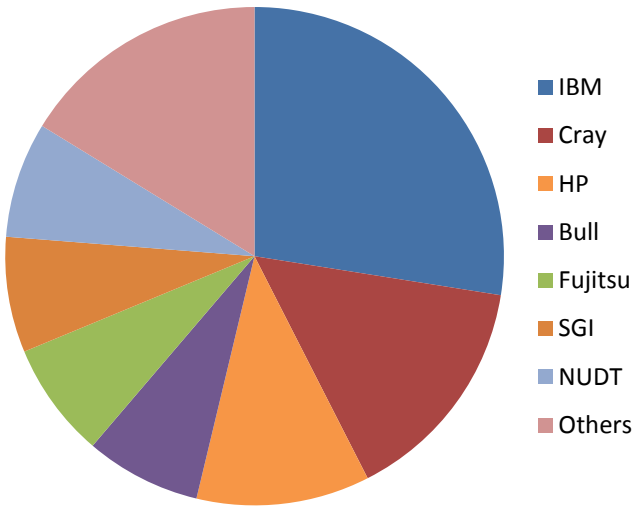

Figure 10: Petascale systems by vendor

IBM is especially focused on its lightweight approach (Blue Gene/Q), which represents 64% of all its systems (and almost 18% of the total), while the rest have a traditional architecture (36%). Cray, on the other hand, prefers the traditional (67%) and accelerated (33%) architectures, with absolutely no models based on a lightweight architecture. HP is exclusively present with accelerated systems. It is interesting to see that, although there is somewhat of a balance between the three architectures in terms of number of petascale systems, vendors don't offer all three but instead tend to be partial towards one.

4.1.3.6 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 more than 57% of the petascale systems. The usual runner-up, AMD, is now behind IBM in the fight for the second most popular processor in petascale computing, with IBM PowerPC processor (available only on Blue Gene/Q systems) taking 18% market share while AMD Opteron's only manage 15%. The remainder of the market consists of the Fujitsu SPARC64, the ShenWei SW1600, and the IBM POWER7.

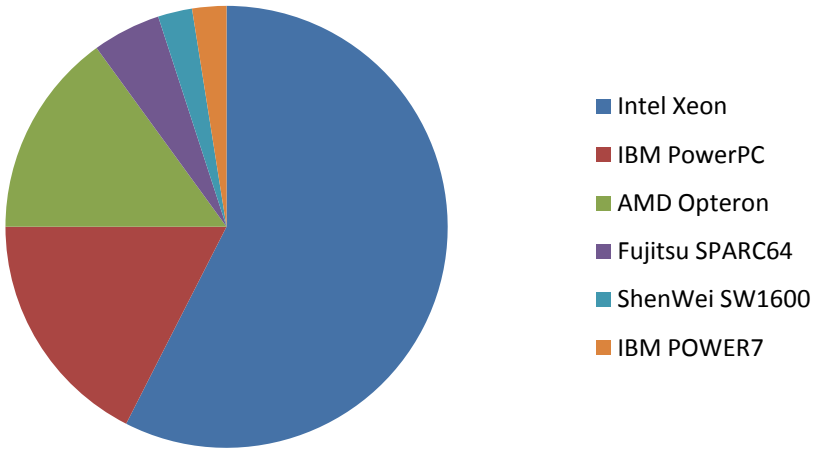

Figure 11: Petascale systems by processor

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 (the speed of IBM's PowerPC processor) and 3 GHz, with 95% falling in this range. The average clock speed for all petascale systems is around 2.3 GHz.

4.1.3.7 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 (62.5%). Of the fifteen systems that do have an accelerator, eleven use some form of NVIDIA Tesla GPGPU. The next most popular accelerator, present in three systems, is the new Intel Xeon Phi coprocessor, based on a more traditional x86 paradigm. The only other accelerator found on the list is the AMD FirePro, a GPU not even specifically tailored for general-purpose computing. 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.

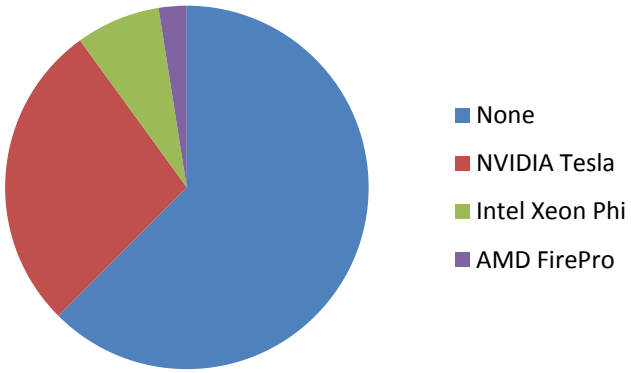

Figure 12: Petascale systems by accelerator

4.1.3.8 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 5k and 385k cores, with an average of 76k and a median of 29k; many-core systems have between

77k and 1.5M cores, with an average of 450k and a median of 280k; and traditional systems have between 50k and 150k, with an average of 100k and a median of 110k. By looking at Figure 8 it is clear that lightweight architectures tend to be at the high end of the core count (80% are above 100k cores), while accelerated systems take the low positions (87% are below 100k cores). Traditional supercomputers lie around the 100k core count line. There was taken into account only the core count of the GPs, not the cores of the accelerators (GPU).

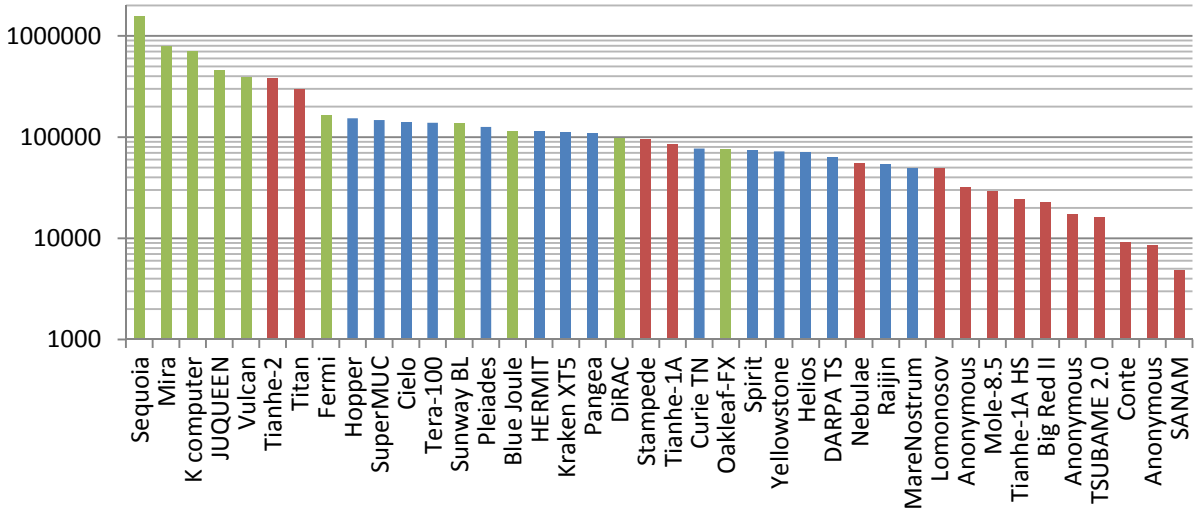

Figure 13: Core count of petascale systems

4.1.3.9 Memory

The amount of memory in each system varies up to two orders of magnitude, from 15 TB to 1.5 PB, with an average of 310 TB and a median of 170 TB. It is very difficult to reach meaningful conclusions comparing the amount of memory for several reasons. For one, the size of the system is obviously an important factor for determining the memory but even so the relation is not closely correlated, and it occurs that a large system such as Titan has less memory than Mira which has almost one third of the performance. Also, memory is an optional field when registering for Top500, so there is no control and several systems do not provide it. This is especially important in the case of accelerated systems, where the co-processor memory is sometimes added to the system memory and other times not, making comparisons very difficult.

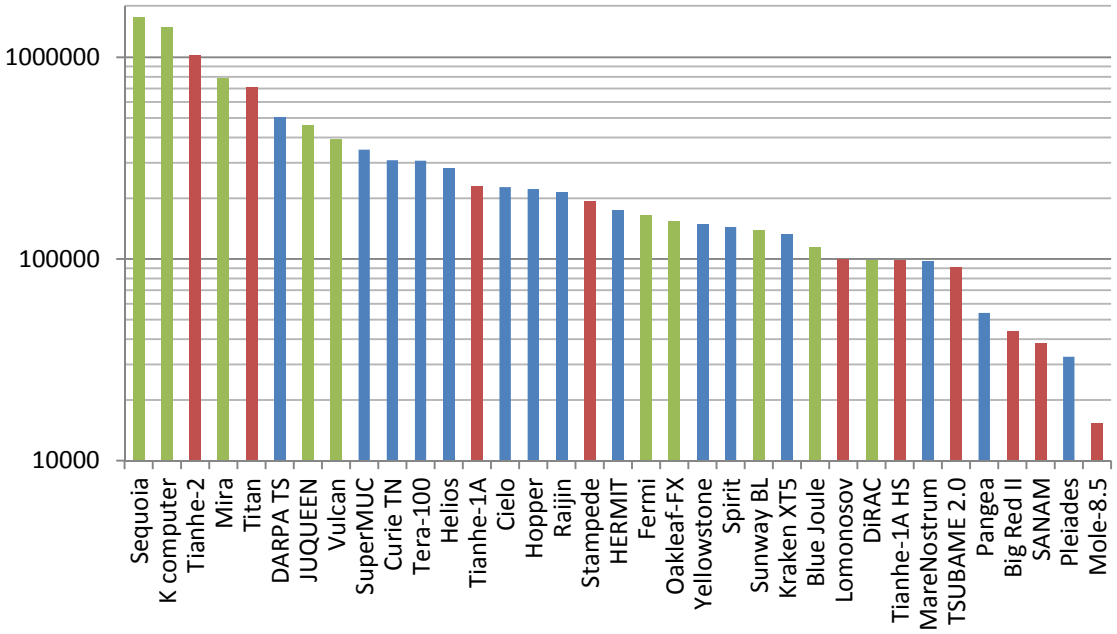

Figure 14: Memory of petascale systems (in TB)

4.1.3.10 Interconnects

A total of eleven different interconnects technologies are used throughout the petascale systems, of which the most common 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.
- **Gigabit Ethernet** – An IEEE standard (802.3), it transmits Ethernet frames at a rate of 1Gbit/s.

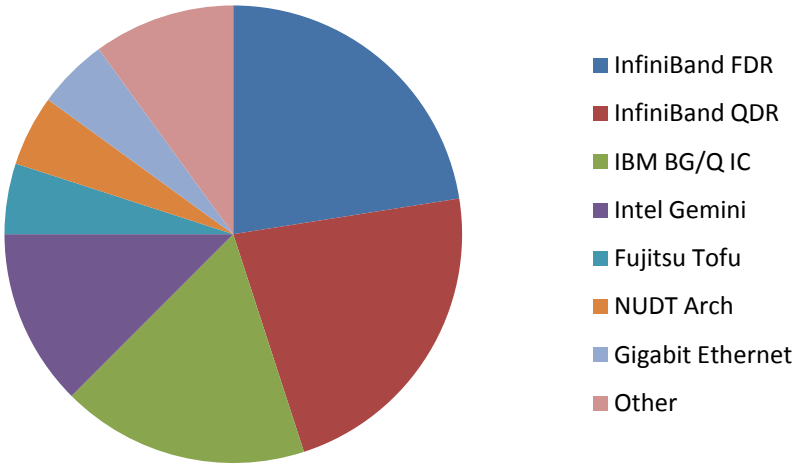

Figure 15: Petascale systems by interconnect

The most popular of these interconnects are the two variants of InfiniBand: the older and slower QDR and the new FDR, with 22.5% market share each. IBM's BG/Q interconnect is used solely on Blue Gene/Q machines, yet still takes almost 18% of the market, followed by Intel Gemini (12.5% share). Fujitsu Tofu, NUDT Arch, and Gigabit Ethernet each have two systems, representing a 5% share. The other interconnects used are the IBM P7 IC (used in the IBM DARPA prototype), Intel SeaStar2+ (originally from Cray), 10G Ethernet (the successor of Gigabit Ethernet), and TH Express-2, the new interconnect designed by NUDT specifically for Tianhe-2 (could be considered the successor of Arch).

4.1.3.11 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 28% and 93%, with an average of around 72%. It is important to note that this ratio is strongly related to the code used to measure "real-world" performance. A system with a 28% efficiency running LINPACK could in theory have 90% efficiency on other benchmarks or on actual applications running on the machine. The same is true for the opposite situation, where efficiencies on LINPACK can be higher than real-world computing efficiencies. It is not possible to make an extrapolation from LIPACK performance to a real application performance with general validity independent from the architecture.

Similarly to core count, computing efficiency is very different depending on the architecture of the system. Accelerated systems average only 52% efficiency, with a maximum of 72%, which is the average for all the systems combined and less than the minimum efficiencies of both traditional and lightweight architectures. Many-core and traditional set-ups are much more similar in terms of efficiency, with many-core slightly ahead (86% efficiency on average, compared to 84% for traditional machines).

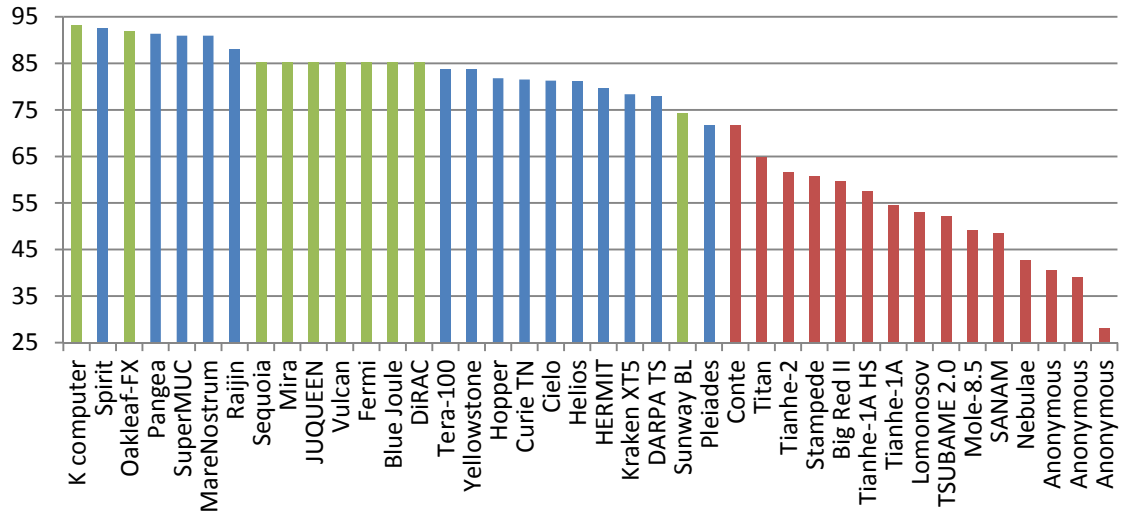

Figure 16: Computing efficiency of petascale systems (in %)

4.1.3.12 Power efficiency

In today's striving for more energy efficient systems, power efficiency imposes as one of the most important metrics. Expressed in MFlop/s/W (ratio between sustained performance of LINPACK execution and the power consumption during the execution) it is used by the Green500 list to provide a ranking of the most energy-efficient supercomputers in the world.

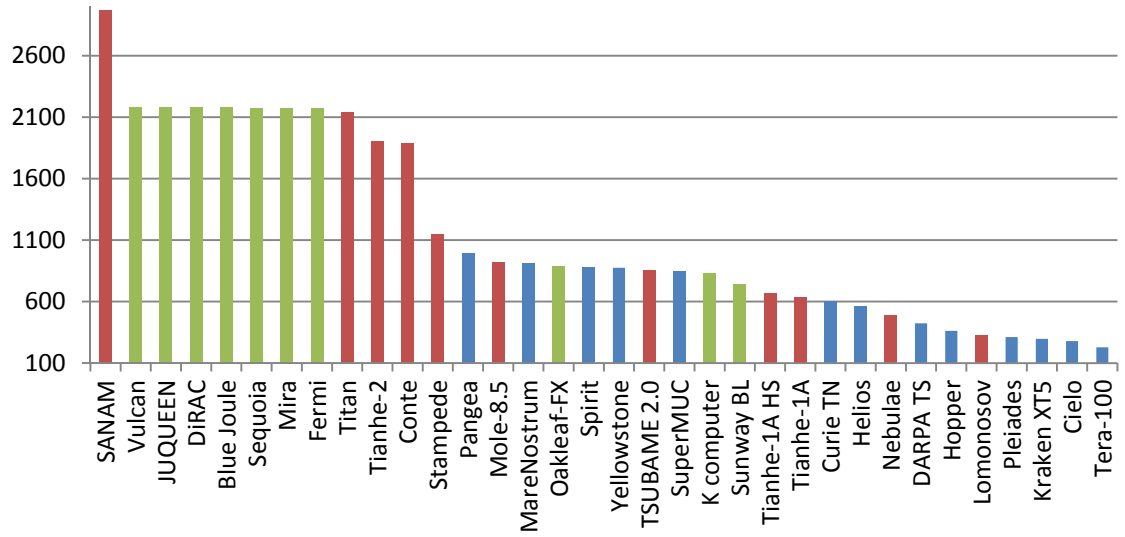

Figure 17: Power efficiency of petascale systems (in MFlop/s/W)

This petascale list clearly distinguishes highly efficient systems which are comprised of IBM Blue Gene/Q systems (seven systems with roughly the same power efficiency) and accelerated systems: SANAM based on AMD FirePro S10000 GPUs with the highest efficiency on the list (close to 3 GFlop/s/W), Titan with NVIDIA GPUs, and two Intel Xeon Phi machines, Tianhe-2 and Conte. The overall average for all petascale systems is 1.15 GFlop/s/W, which can be decomposed by architecture as: 1.3 GFlop/s/W for accelerated systems, 1.8 GFlop/s/W for many-core systems, and 585 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 more than double and triple the average efficiency, respectively.

4.1.4 Dynamic Analysis

Having an overview of the current situation of the world-class HPC market is useful, but it is also much more interesting having this general view over the time. 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.

4.1.4.1 Number of petascale systems

The number of petascale systems in the world has been practically doubling each year for the past 5 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.

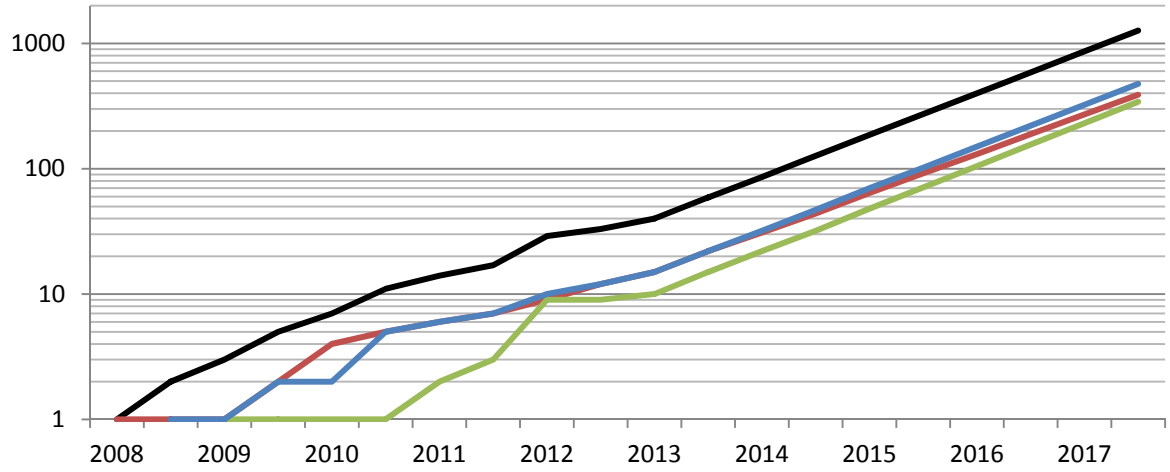

Figure 18: Evolution and prediction (from 2013 onwards) of the number of petascale systems total (in black), broken down by architecture: accelerated (in red), lightweight (in green), and traditional (in blue)

From the point of view of the hardware architecture the market has been evolving with great sways up until now, which makes forecasts ever more uncertain, but the general trends seem to show that all three techniques (accelerated, lightweight, and traditional) are growing, but with different speed. Of course this must only be taken as rough trends and projections as of today, because the number of points and smoothness of curves are rather limited and unequal to allow reliable quantitative extrapolations.We only assume the techniques will still remain on the market in the near future.

4.1.4.2 Year of construction

As we have already seen, the number of petascale systems doubles annually, which means new systems take around 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).

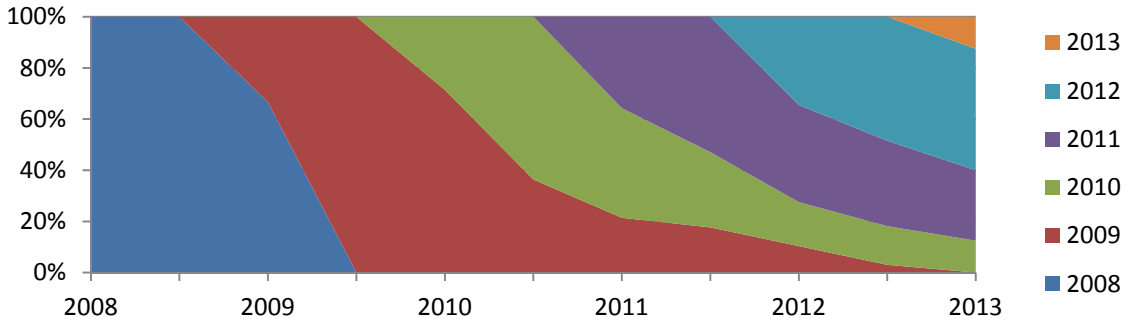

Figure 19: Evolution of the market share for deployment year of petascale systems

4.1.4.3 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, an achievement that they have repeated now in 2013.

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.

Germany was the second country to have a petascale system on the Top500 list in 2009, and are still available in 2013 in the top10 list with 2 systems (FZJ, LRZ).France joined the petascale race in 2010, while the UK, Italy, and Spain have entered more recently. Together, the four European nations make up 25% of the market share, which would in fact be second place between China and the USA.

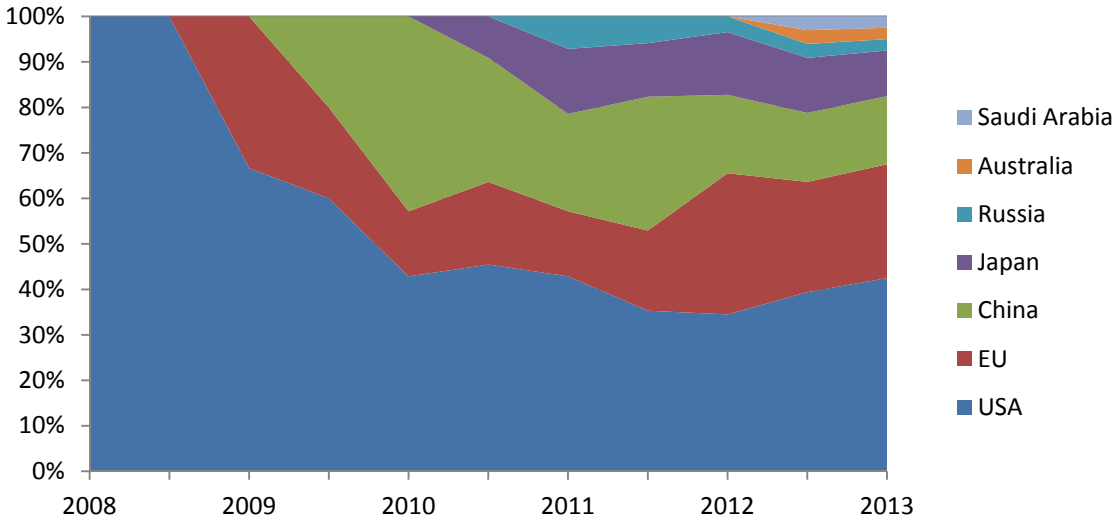

Figure 20: Evolution of the country of petascale systems

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

4.1.4.4 Performance

The current rate of growth for maximum performance (both peak and LINPACK) is around 1.5x annually. If this trend continues, 100 PFlop/s systems should be available as soon as 2014, and the first exascale machine (in peak performance) will appear sometime between 2016 and 2017.

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 Tianhe-2's (62%), 1 EFlop/s in LINPACK would require more than 1.6 EFlop/s peak, which this model places in the 2017-2018 timeframe. However most of the experts doubt that this will happen in 2018. The 2019 is the current milestone in their opinion, due to many different exascale issues (see chapter 6).

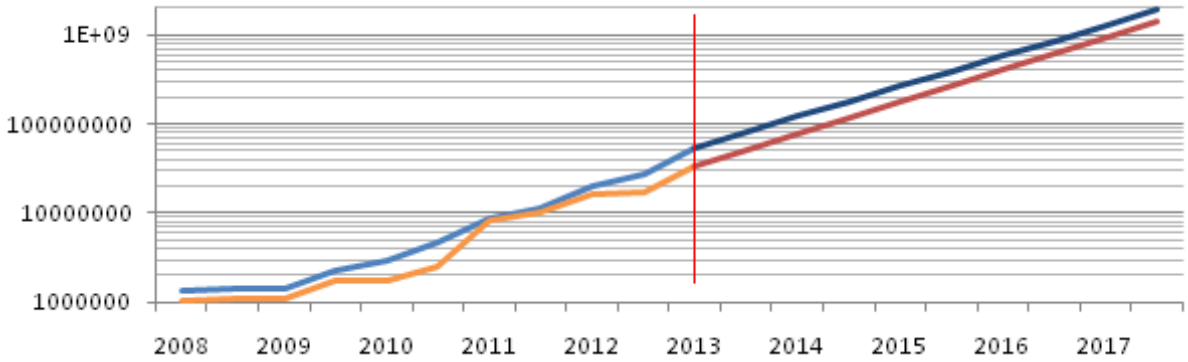

Figure 21: Evolution of maximum LINPACK (orange) and peak (blue) performance (with predictions starting from mid of 2013 – red line)

4.1.4.5 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.

Hewlett-Packard participated with NEC in one of the first batch of petascale supercomputers back in 2010, but didn't create their own petascale system until this year, when three pure-HP systems were added to the list, all three of them built for commercial purposes. This is the typical behaviour of HP, which doesn't usually rush to make Top50 systems, but joins in when the market is more open and lucrative. SGI has had a very similar evolution, from only one system in 2011 to 3 systems now in 2013.

Bull and Fujitsu presented their first petascale machines in 2010 and 2011, respectively, and since then have added a few new systems periodically, but not enough to improve or even

keep their market share. Since NUDT is not a commercial vendor but an experimental institution, it is logical that they are not striving to keep any market share, but instead have been releasing a high-end computer every 2 years.

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 HP, SGI, Bull, and Fujitsu (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.

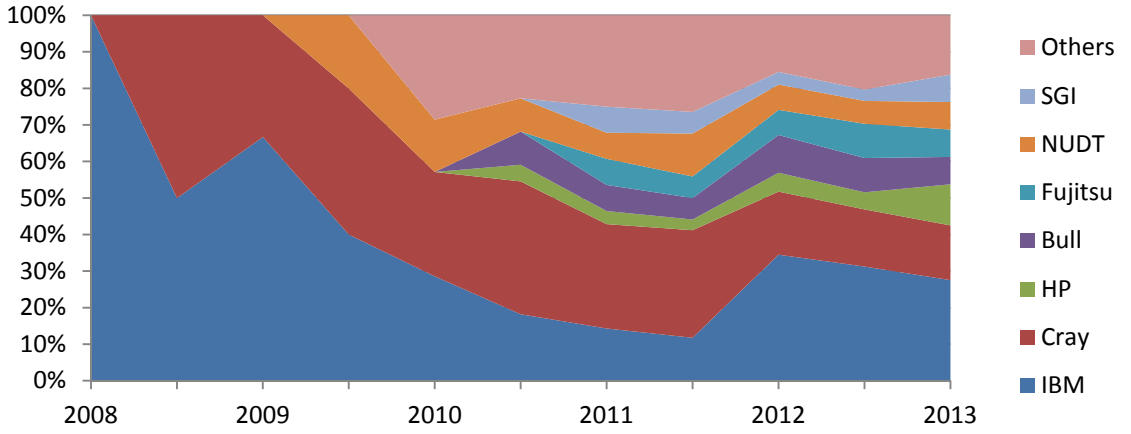

Figure 22: Evolution of vendors of petascale systems

4.1.4.6 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, and now they have completely turned the tide with almost 60% of the market.

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). IBM also has a POWER7-based petascale system, which might help them add a little more to their market share in the future, although the future of this processor is quite unclear.

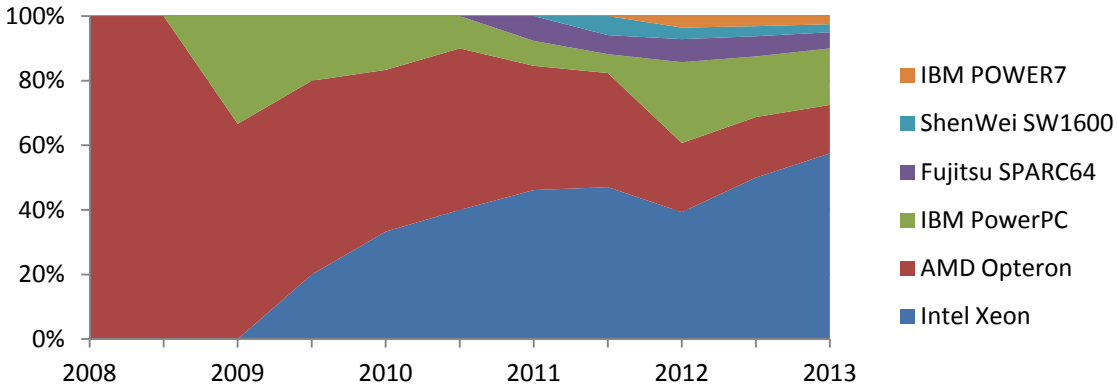

Figure 23: Evolution of processors used in petascale systems

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). Little has changed in the past 2 years with respect to these two processors, but there are other new lines rumoured to appear in the following years.

4.1.4.7 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.

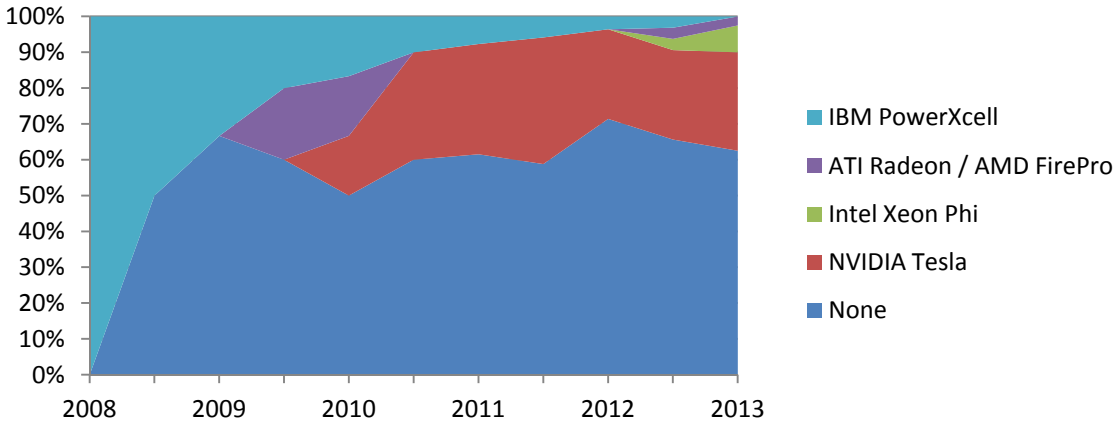

Figure 24: Evolution of accelerators used in petascale systems

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 originally in Stampede and then in two more petascale systems including the leading Tianhe-2 computer, is based on a traditional x86 microarchitecture with stream processing. This specifically designed co-processor (not based on GPU) is also quite energy efficient.

4.1.4.8 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 Blue Gene/P and Blue Gene/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 IBM Blue Gene IC variants on the other hand have seen their share fall constantly after its first generation Blue Gene/P supercomputer model until the introduction of the next BlueGene/Q model in 2012, where they received huge boost. 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. The newcomers in 2013 are Gigabit Ethernet and 10G Ethernet, two more industry standards to compete with InfiniBand (although probably only on the low-end because of their higher latency).

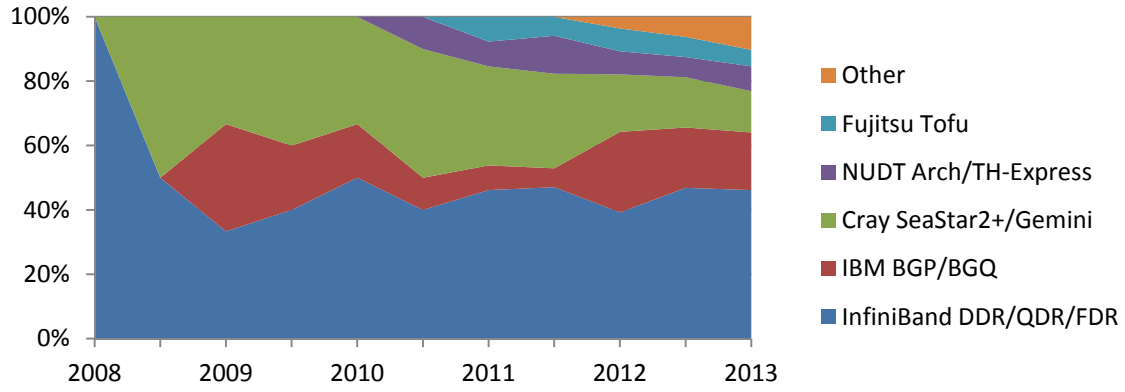

Figure 25: Evolution of interconnects used in petascale systems

4.1.4.9 LINPACK Efficiency

With regards to LINPACK execution, the efficiency of petascale systems has seen both a 19% rise in its maximum and a 47% 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 70-75%, and the median has remained between 75% and 80%.

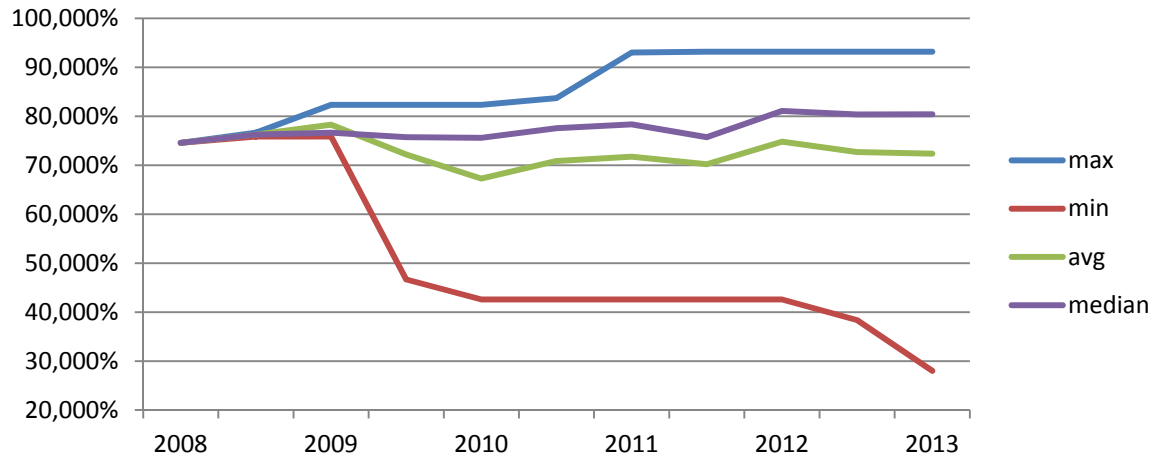

Figure 26: Evolution of the computing efficiency of petascale systems (in %)

This large discrepancy between the maximum and minimum values for computing efficiency shows that there is an underlying problem with using the LINPACK benchmark for calculating real-world performance. This subject is covered in more detail in the following chapter: Beyond Top500.

4.1.4.10 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 2018 and 2019, which is actually a later timeframe than that seen earlier based purely on performance, indicating that although it might be theoretically possible to build an exascale machine in 2017, it won't be power-effective until 2018. The question then is whether the 20MW limit will stand, or if the desire for exascale will be enough to warrant a higher energy envelope.

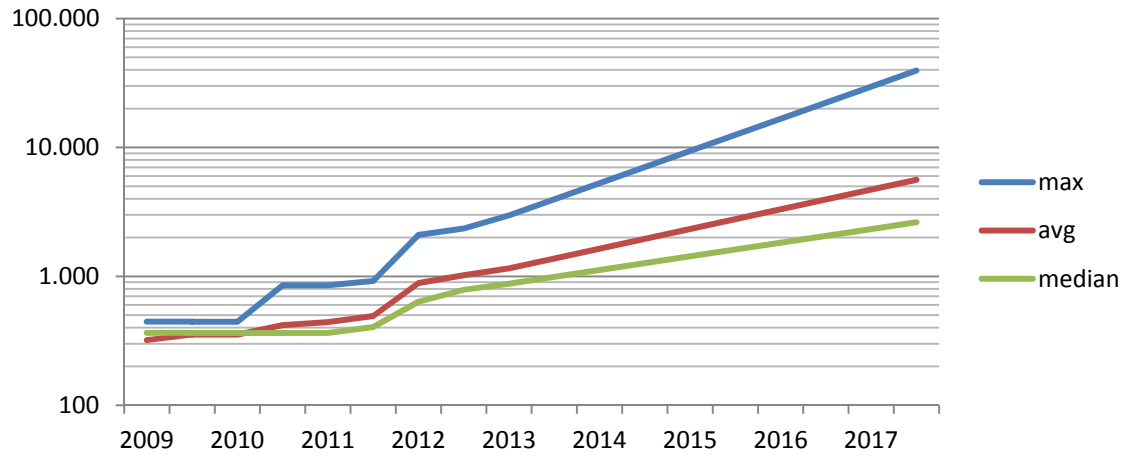

Figure 27: Evolution and prediction (from 2013 onwards) for power efficiency of petascale systems (in MFlop/s/W)

4.1.5 Beyond Top500

In the last couple of years, there has been a lot of criticism on the Top500 list and its relevance and impact to HPC, and that topic was also tackled by various talks at ISC'13 [20]. It is generally accepted that Top500 is an important cornerstone of HPC and that it has tremendous value because twice a year it gives a reliable snapshot of the recent situation of supercomputing. Additionally and maybe even more important, it provides the wealth of statistical material collected through years that allows deeper view into the technical and organizational history and trends in supercomputing.

The main criticism of the list is that it is based on only one benchmark – High Performance Linpack (HPL) which becomes more and more unreliable as a true measure of system performance for a growing collection of important science and engineering applications. According to Heroux and Dongarra [21], without some intervention, future architectures targeted toward good HPL performance will not be a good match for real applications.

Unfortunately, currently it seems that there is no other suitable benchmark, as all of them cover only a small part of applications. Recently, there are steps in direction to define a new

high performance conjugate gradient (HPCG) benchmark (Heroux and Dongarra), which will satisfy two important requirements: to accurately predict system rankings for the target suite of applications (ranking of computer systems using the new metric must correlate strongly to how real applications would rank these same systems), and to drive improvements to computer systems to benefit users applications (when metric results are optimized for a particular platform, the changes will also lead to better performance in real applications). They do not propose elimination of HPL as a metric, since historical importance and community outreach value of HPL (and current Top500 list) is too important to be abandoned. Instead, proposed HPCG benchmark could serve as an alternative ranking of the Top500 list (similar to the Green500 re-ranking of the items on this list).

An additional effort in complementing Top500 list is the Graph 500 list [22]. As current benchmarks and performance metrics do not provide useful information on the suitability of supercomputing systems for data intensive applications, a new set of benchmarks is needed in order to guide the design of hardware architectures and software systems intended to support such applications. This list is the first serious approach to complement the Top500 with data intensive applications. It ranks the world's most powerful computer systems for data-intensive computing and gets its name from graph-type problems which are at the core of many analytics workloads in applications. It is backed by a steering committee of over 50 international HPC experts from academia, industry, and national laboratories. The Graph 500 benchmark is based on a breadth-first search in a large undirected graph, and in contrast to the Top500 list, which uses TFlop/s as a metric, Graph 500 uses GTEPS (billions of traversed edges per second).

As energy consumption becomes a limiting factor, making big data computing energy-aware is an imperative. In a similar way that Top500 list is accompanied by Green500 list, Graph 500 list is accompanied by Green Graph 500 list [23]. The data on this list is collected in collaboration with the Graph 500 list, and the benchmark and the performance metrics are identical with Graph 500. It is also designed to complement the Green500 list with an energy metric for data intensive computing, and in order to allow comparisons it strives to keep the differences in the rules to Green500 low.

During ISC'13 a new Graph 500 list for June 2013 was published (it follows the publications of Top500 lists). The first official Green Graph 500 list (for June 2013) was also announced, and both lists are dominated by IBM Blue Gene/Q systems.

4.2 Business Analysis

This chapter provides information on several topics regarding the HPC market in general from a business intelligence perspective based on information gathered at the last International Supercomputing Conference (ISC 2013 in Leipzig, Germany).

4.2.1 Current buzzwords

Heterogeneity and neo-heterogeneity were buzzwords often repeated by Intel at ISC'13. Neo-heterogeneity is defined as a heterogeneous system with a single programming model, where the hardware architecture has multiple classes of compute capabilities that are accessed by a common programming model, streamlining development and optimization processes – an advantage not possible when using a combination of CPUs and GPU accelerators. Intel was using these buzzwords while strongly promoting their Xeon Phi products and the newest No.1 on the Top500 list, Tianhe-2, was their most important showcase. Tianhe-2 (Milky Way 2) uses 48,000 Xeon Phi accelerators coupled with 32,000 Ivy Bridge Xeon CPUs to achieve 33.86 PFlop/s.

4.2.2 Memory

During the ISC'13 important topics covered by several lectures were current state and future developments of DRAM memory technologies [24][25]. DRAM is not scaling with Moore's Law and the gap between the DRAM performance improvement rate and the processor data consumption rate is continuously growing. Besides the upcoming evolutionary DDR4 memory, Hybrid Memory Cube (HMC) [26] was often mentioned as a solution (at least a temporary one) in the DRAM revolutionary context.

At the core of the HMC is a small, high-speed logic layer that sits below vertical stacks of DRAM die that are connected using through-silicon-via (TSV) interconnects. HMC should deliver significant improvements in performance (a single HMC should provide more than 15x the performance of a DDR3 module), energy efficiency (utilizing 70% less energy per bit than DDR3 DRAM technologies) using nearly 90% less space than today's RDIMMs. The HMC Consortium (HMCC), a working group made up of industry leaders that might build, design, or enable HMC technology, has defined the HMC interface specification.

4.2.3 Storage

The Memory Channel Storage (MCS), which uses the standard RDIMM memory interface to support the lowest latency for SSD storage, is attached via the DDR-3 CPU channel. This technology represents a drastic improvement in latency reduction, which ranges from 85% compared with PCIe based SSD storage to as great as a 96% reduction when compared to SATA/SAS based SSD storage. In addition to block storage, MCS can also enable system memory to expand from gigabytes to terabytes.

Diablo's MCS allows flash memory to be used for the first time in ultra-low latency applications that demand deterministic latency with no performance load spikes. This technology is also extremely valuable for blade servers, where the size and number of PCIe slots limits the range of scalability via traditional methods of adding SSD storage. Putting the SSD storage directly on the CPU memory bus allows the entire application data set to reside in CPU memory space which gives another magnitude of performance to demanding enterprise applications [38].

4.2.4 Intel or HPC accelerator

More than 80 percent (403 systems) of the supercomputers on the 41st edition of the Top500 list are powered by Intel processors, and out of those systems that are making their first appearance on the list, Intel-powered installations account for 98 percent.

One of the biggest announces of the ISC'13 was expansion of Intel's Xeon Phi coprocessors[24] portfolio and revealing of details of the second generation of Intel Xeon Phi products code named "Knights Landing". The announced Intel Xeon Phi products were:

- The top end Intel Xeon Phi coprocessor 7100 family with the highest level of features, including 61 cores clocked at 1.23GHz, 16 GB of memory capacity support (double the amount previously available in accelerators or coprocessors) and over 1.2 TFlop/s of double precision performance,
- The Intel Xeon Phi coprocessor 3100 family that tries to balance performance with the cost. The family features 57 cores clocked at 1.1 GHz with 6GB of RAM and 1 TFlop/s of double precision performance,
- Intel Xeon Phi coprocessor 5120D, another product to the Intel Xeon Phi coprocessor 5100 family announced last year, that is optimized for high-density environments with the ability to allow sockets to attach directly to a mini-board for use in blade form factors.

Codenamed "Knights Landing," the next generation of Intel MIC architecture-based products will be available as a coprocessor or a host processor (CPU) and manufactured using Intel's 14nm process technology featuring second generation 3-D tri-gate transistors. As a PCIe card-based coprocessor, "Knights Landing" will handle offload workloads from the system's Intel Xeon processors, but as a host processor directly installed in the motherboard socket, it will function as a CPU (handling simultaneously all the duties of the primary processor and the specialized coprocessor). Also, when used as a CPU, "Knights Landing" will remove programming complexities of data transfer over PCIe, common in accelerators today. In addition, according to Intel, memory bandwidth for all "Knights Landing" products will be significantly increased by introducing integrated on-package memory that should further boost the performance for HPC workloads.

The June edition of the Top500 list had recorded 11 systems based on the Intel Xeon Phi coprocessor, including the Petaflop/s class systems like Milky Way 2 at 54.9 PFlop/s and Stampede at 8.5 PFlop/s of peak performance.

4.2.5 Large Systems Vendors

This is a list of vendors which were reviewed by the PRACE team at the ISC2013 conference and exhibition. Although the list is not exhaustive, it gives a good reflection of new trends in HPC.

4.2.5.1 Bull

New supercomputing systems that have been delivered by Bull:

- The first stage of the bullx supercomputer ordered by the French weather forecasting agency has been installed at Météo-France's site in Toulouse and will be fully available for production this year. This supercomputer is equipped with the upcoming Intel Xeon E5-2600 v2 processors, based on Ivy Bridge microarchitecture and 22nm manufacturing process. It is also the largest system to rely on the direct liquid cooling technology developed by Bull to lower energy consumption (using bullx DLC B710 compute nodes).
- Cartesius, the new national supercomputer at SURFsara, Netherlands, which will exceed 1 PFlop/s by the end of all installation phases.

Two main HPC server lines for Bull are air cooled B500 blade system and direct water cooled B700 DLC blade system, where cabinets support up to 80kW of electrical power. Both of these lines support regular (B510 and B710) and accelerated blades (B515 and B715) supporting both NVIDIA Tesla GPUs and Intel Xeon Phi coprocessors. Within the bullx product line for HPC, Bull also offers SMP bullx supernodes with support for large amounts of memory and high expandability (up to 4 x 4-socket SMP nodes can be interconnected through a Bull-designed switch, to form a large SMP node - with up to 16 sockets and 4 TB of memory) and bullx rack-mounted series (family of 2-socket and 4-socket servers, based on the latest generation of Intel Xeon or AMD Opteron processors).

Bull has inaugurated its Center for Excellence in Parallel Programming in Grenoble, France. Its mission will be to support engineers and scientists in research centers and industry to overcome the critical technological barrier of HPC application parallelization. This center will benefit from close cooperation with Intel in the area of supercomputers and parallel computing and partners that are involved in the Center already include Allinea, CAPS and the Joseph Fourier University.

4.2.5.2 CRAY

CRAY's leading products are its new XC30 systems (previously codenamed "Cascade,"), which are available in two versions:

- XC30 with transverse air-flow liquid-cooled architecture,
- XC30-AC - air-cooled version with smaller and less dense supercomputing cabinets with no requirement for liquid coolants or extra blower cabinets.

XC30 systems support CRAY's Aries Interconnect and Dragonfly Topology and Intel Xeon processors.

CRAY also offers Big Data solutions: YarcData Urika Big Data Graph Appliance and Cray Cluster Supercomputers for Hadoop.

Urika system is built for graph analytics with large, global shared memory whose architecture can scale up to 512 terabytes, massively multithreaded graph processor Threadstorm that supports 128 hardware threads in a single processor (65,000 threads in a 512 processor system and over 1 million threads at the maximum system size of 8,192 processors) and highly scalable I/O.

Cray Cluster Supercomputers for Hadoop is based on Cray CS300 cluster supercomputer series with Intel Distribution for Apache Hadoop software.

4.2.5.3 NEC

NEC is still working on their Next Generation Vector machine (previously known as SX-X), successor of SX-9 system. This system is announced for 2014 and it will use one 4-core processor per node with a performance of 256 GFlop/s and 256 GB/s memory bandwidth (1 byte per flop ratio). The aim of NEC is to have ten times better performance per kW in SX-X than it predecessor SX9.

4.2.5.4 Eurotech

Eurotech is a global company with a strong international focus: born and still headquartered in Italy, it has operating locations in Europe, North America and Japan. The company is grounded in the field of embedded computers and real-time control of machines and is focusing the HPC market since the birth of the AURORA platform few years ago. Its systems are based on the x86 architecture, integrating processors from Intel, and accelerators from NVIDIA and Intel (MIC). A few key aspects in AURORA systems are: High performance density with liquid cooling, that allows reaching very high densities: the hybrid Aurora Tigon can pack 1.3 Petaflop/s in 5 m^2. The energy efficiency is achieved with hot liquid cooling, to leverage free cooling even at hot climate latitudes. Very efficient power conversion, having racks ready for AC/DC conversion outside. Aurora's direct hot water cooling technology, according to Eurotech, should work with an inlet water temperature of above 50 °C enabling high density packaging up to 100 kW per rack.

In fact, the system named Eurora-CINECA, an AURORA prototype with NVIDIA K20x accelerators, gained the first position in Green500 list on June 2013 (3.21 gigaflop/s/watt), while the second position in list is for another Italian AURORA installation. On the side of scalability the AURORA Tigon scales linearly with a 3D Torus and limits OS jitter thanks to

an independent synchronization network. There is still an Infiniband network that can coexist with the optional FPGA driven 3D Torus.

Eurotech also offers the Aurora G-Station, advertised as a supercomputer in a box, deployable with minimal infrastructure. Within the small package it can deliver 21 TFlop/s. The G-Station contains one Aurora HPC 20-22 chassis with 8 slots and each chassis provides electrical, network (40Gbps QDR IB) and liquid connections and mounts up to 8 Aurora HPC 20-23 blades. All computational nodes in the Aurora G-Station are water cooled by using Aurora Direct Hot Liquid Cooling. G-Station comes in two variants:

- Split, with computational unit and external cooling unit,
- Standalone, with embedded liquid cooling (slightly larger packaging compared to the split system).

It supports up to 16 Intel Xeon E5 processors and up to 16 Nvidia Kepler K20 or Intel Xeon Phi accelerators.

4.2.6 New CPU architectures

This section will focus on the non-x86 market, and the alternatives to the current standard architecture of HPC systems. In most cases the architectures are not new per se, but rather the usage of them in a HPC context. One of the solutions is PowerPC, mentioned recently as a very successful architecture. It is not listed here, because there is no information on new developments.

4.2.6.1 ARM

Overview

For several years now the ARM architecture has been up and coming, yet it has never featured on the Top500 list. Being developed by the UK based ARM Ltd, it is of course of special interest to the European audience. A large number of silicon vendors are licensing the ARM IP core from ARM Ltd, but mostly it is being used in SoCs for the mobile market and also embedded in other products. The following will focus on the chip designs targeted towards the server market, with support for ECC DRAM as example of a server feature.

Given that ARM only licenses technology and IP blocks, but does not produce CPUs itself, there are a wide variety of features in ARM designs from different vendors. You can license the building blocks for an entire SoC from ARM, but many vendors use their own IP blocks outside the CPU core. Differentiating for example on the interconnect is a way for vendors of ARM based chips to compete with each other.

AArch64 is the official name of the 64-bit version of the ARM architecture, and is also known as ARMv8. It is a redesign of the ARM architecture and switches to what is now known as the AArch32 execution state when running 32-bit code. Among the new features of ARMv8 are double precision SIMD instructions and 32 128-bit registers. The ARM designed SoC IP blocks (“System IP”) includes a memory controller with DDR3, DDR4 and ECC support in the 500 series.

Two application processors, the performance oriented A57 and the low power A53, will be the first two AArch64 cores. Systems for the HPC market will probably be using the A57 core.

Mali is a GPU designed by ARM Ltd. and included in many ARM designs for the embedded market. It is not however used in the ARM designs targeting the compute market, and it has not had OpenCL support until the T604 model. According to the ARM Mali roadmap the 600

family will be expanded with more general purpose and higher performance designs in the near future. For the time being the ARM systems with GPUs are Nvidia based, though.

Red Hat Enterprise Linux (RHEL) and SuSE Linux Enterprise Server (SLES) and derivatives of them have been widely used in x86 HPC deployments. For ARM based systems many vendors are providing Ubuntu based software stacks, so this may mean that the software landscape will look different in the future. Partly, this is of course due to the fact that neither RHEL nor SLES exists in ARM versions, but also due to Canonicals participation in Linaro, which has based its reference distribution on Ubuntu. Whatever distribution will be used it will certainly be Linux based, so no major changes are to be expected compared to x86.

ARM chip vendors

As noted above, there are many more companies selling ARM based CPUs, this table only includes designs used in servers. For the 64-bit column, only product announcements have been used, not a general licensing of AArch64 technology.

Vendor	ARM 32-bit	ARM 64-bit (announced)
AMD	No	Yes, 2014
Applied Micro	No	Yes
Calxeda	Yes	Yes, 2014
Cavium	Yes	Yes
Marvell	Yes	
Nvidia	Yes	
Samsung	Yes	
STMicroelectronics	Yes	Yes
Texas Instruments	Yes	

Table 4 List of ARM chip vendors

AMD

During 2012 AMD announced that they had become an ARM licensee and would be producing ARM based Opterons with the first products appearing in 2014. In June 2013, a more concrete roadmap was shown with an A57-based “Seattle”. This would also include both 10 Gbit Ethernet and the SeaMicro Freedom Fabric interconnect integrated into the chip.

Applied Micro

Having previously produced a number of processors based on the Power architecture for the embedded market, Applied Micro announced that its first ARM design would be a 64-bit one. Known as X-Gene, it has clearly aimed at being one of the first 64-bit implementations to market with FPGA versions being shown as early as 2011 when ARMv8 was announced. A silicon implementation was first shown publicly running Linux in June 2013.

Calxeda

Has been developing ARM based processors firmly targeted at the server market for several years, and has been focusing on their own fabric interconnecting SoCs for both clustering and

management. Apart from the interconnect Calxeda has been focusing on the energy consumption, and markets their current processors as using 5W per node, DRAM included.

Nvidia

Tegra is a series of mobile processors that were used in an earlier Mont-Blanc prototype. Combining an ARM core with Nvidia GPU core is the strategy taken by Nvidia, meaning that GPU based systems could be self hosting with the ARM cores replacing the x86 host. CARMA is the development kit for this CUDA on ARM approach. No firm dates have been given for an ARMv8 based version of Tegra, but recent roadmaps seem to point towards 2015.

Samsung

Current ARM products from Samsung include the Exynos series, used by the Mont-Blanc project. However, the Exynos family is intended for the mobile market, and it lacks ECC support for example. It does have a Mali T604, which will be used for computation in the Mont-Blanc system.

STMicroelectronics

Long background using ARM cores in its microcontrollers and ASICs built for other companies, also ARM based mobile chips via ST-Ericsson. ST is developing an ARMv8 server class reference implementation.

Of European interest are the manufacturing capabilities of ST. The ST Crolles fab has a 28 nm FD-SOI pre-production process, expected volume production by the end of 2013. Second source of this process technology at the Global Foundries Dresden fab in 2014.

Texas Instruments

TI is covered in the DSP section, since their current ARM products aimed at the HPC market include DSP cores. They also have other ARM products, but these are intended for embedded usage.

ARM server system vendors

The ARM server market is still not as mature as the x86 market, and many of the entrants are still on the level of pre-production and only available to selected customers.

Vendor	Product	CPU vendor/model	Shipping (GA)
Boston	Viridis	Calxeda ECX-1000	Yes
Dell	Copper	Marvell Armada XP	No
	Iron	Applied Micro X-Gene	No
E4 Computer Engineering	Arka	Nvidia Tegra 3	Yes
HP	Moonshot	Applied Micro X-Gene	No
		Calxeda ECX-1000	No
		Cavium “Thunder”	No

Table 5 List of ARM server system vendors

Boston Viridis

UK based company that uses the Calxeda ECX-1000 in either a 2U compute chassis or a 4U storage chassis. Viridis is based on the Calxeda EnergyCard, which bundles four nodes on each card, and the server will host up to 12 cards for a total of 28 nodes. Each node has 4 GB RAM, which is consistent with the 32-bit ECX-1000.

E4 Arka

Combining the Nvidia Tegra 3 with the Nvidia Quadro GPU yields the Arka system from Italian company E4 Computer Engineering. The hardware is designed by another Italian company, SECO, which also did the Kayla development kit for Nvidia. The system used in the Arka system seems to be a more commercialized version of this development kit.

It is sold in different configurations, including both blades and a prepackaged microcluster. The last one also includes a x86 based management node, so it is not a completely ARM based system.

4.2.6.2 DSP

Overview

Digital Signal Processors are currently mainly used in the telecommunication sector, and highly optimized for streaming application. Use of them for HPC applications has been proposed for some time, but both the lack of double precision and the differing programming environment has been hindering DSP uptake.

Texas Instruments

Texas Instruments is not the only company producing DSPs, but is the most visible one in the HPC market. In recent years it has been steadily improving both the hardware, with support for double precision in the Keystone architecture, and the software stack to be able to compete in the HPC market.

For the TI Keystone architecture, a Linux port for C66x cores has been made available, but the hardware lacks full MMU support. This makes it hard to use it in a shared cluster with possibly untrusted users.

In the last few years TI has been combining ARM and DSP cores on the same die on some models, opening up the possibility of using the ARM side as a frontend running a standard Linux based OS. Current KeyStone II SoCs bundle A15 ARM cores with C66x cores.

Software wise the DSPs are treated as accelerators and code using OpenMP and OpenCL will offload computation to the DSP cores. The ARM cores in the chip will handle MPI communication. Since the ARM and DSP cores are sharing the DRAM, it is more comparable to an AMD APU than to a GPU connected via PCI Express.

DSP server system vendors

There are few vendors creating HPC systems based on DSPs, and they are not generally available for purchase yet.

Vendor	Product	DSP Model	Shipping (GA)
HP	Moonshot	Keystone	No
nCore	BrownDwarf	Keystone I and II	No

Table 6: List of DSP server system vendors

4.2.6.3 nCore

The American company nCore has together with the Dutch company Prodrive created a more or less turn-key system named BrownDwarf, based on the TI Keystone architecture. The system was announced at ISC'13 and has currently limited availability. Prodrive has developed the carrier boards in the system, which uses the AMC form factor. This is a standard form factor in the telecom sector, and has many chassis already available. The interconnect is based on Serial RapidIO, which is a standard feature of TI DSP products.

DSP boards are often RAM limited; the AMC modules used in BrownDwarf can take up to 26 GB DDR3 RAM, making them more suitable for HPC nodes.

4.2.6.4 FPGA

No vendors exhibiting FPGAs as a general purpose solution were present at ISC, and there were no presentations on FPGA based HPC computations.

4.2.6.5 Many-core architectures

A number of different architectures have been proposed over the years that share the property of having a large number of cores.

Epiphany

A recent entry in this class of architecture is Parallella, developed by an American company named Adapteva. The system is based on Epiphany co-processors coordinated by ZYNQ70xx dual-core ARM Cortex A9 processors from Xilinx. The Epiphany co-processors are based on a RISC architecture with only 35 instructions and combine one arithmetic-logic unit and one floating point unit, 32 KB static RAM and a router with 4 ports that can reach a 64x64 array of cores for a maximum of 4096 cores. The memory model of Epiphany co-processors allows any core to address the SRAM on any other core since they support a single address space. A DMA is also supported, enabling fast I/O. The currently shipping Epiphany-III cores run at 800 MHz and deliver 51 Gflop/s per watt performance, while the future Epiphany-IV design should scale to 64 cores at 1 GHz and achieve 70 Gflop/s per watt, using only 25 milliwatts per core. In future the roadmap of the company calls for adoption of 7nm manufacturing process around 2017, which will enable development of boards with 1000 cores/chip reaching 2 Tflop/s with 2 watts and later 64K cores/chip reaching 100 Tflop/s with 100 watts. The model of funding for their research that Adapteva used – offering Epiphany boards to HPC enthusiasts through Kickstarter (a crowd funding site for creative projects) at prices as low as 100$, is an interesting approach to widening the developer's adoption of their boards that can have positive impact on the overall HPC space.

Several of the developers have worked on DSP architecture at Analog Devices before. What makes Parallella stand out is the unusual funding, it is partly funded via Kickstarter. It raised almost 900 000 USD during the autumn of 2012 from 4965 backers, evidently finding many individuals both interested in parallel programming and hardware architectures. (The reward for backing the project was mainly early access to the finished hardware.) Hardware design files are also openly available on Github, licensed under the GPL.

Hardware wise it is a Xilinx Zynq 7010 or 7020, with a combination of ARM cores and a FPGA, that acts as a front end to the Epiphany cores. The Parallella boards will have either 16 or 64 Epiphany cores.

As usual, one is tempted to say, this system will be running Linux with a full Ubuntu distribution already shown running. Work on OpenCL and MPI support has been started, but seems tobe in an early stage.

4.2.6.6 SPARC

Currently there are four SPARC64 based clusters on the Top500 list, all of these are Fujitsu systems located in Japan. The most well known of these is the K computer, which helds the top position on two Top500 lists. Fujitsu has developed the "fx" (or HPC) versions of their SPARC64 series, with the IXfx currently being the top model. There is a SPARC64 X available for commercial workloads already that includes the HPC-ACE instructions and on-chip interconnect, and it called a SoC by Fujitsu.

The current number one on the Top500 is based on Intel processors for the compute elements. It does have a SPARC based front end, which uses the Chinese Galaxy FT-1500 CPU. On its own, the front end system would probably be among the top 100 systems considering its estimated peak performance of more than 600 TFlop/s.

Oracle is not present in the HPC market currently, so this leaves the SPARC market to Asian implementations. Little information about especially the Chinese SPARC developments is available, so lots of question marks about future developments remain here.

4.2.7 Industry Segment Systems in the Top500.

This section provides information on the June 2013 Top500 systems that are owned and operated by the industry. It tries to identify the existence and allotment of industrial HPC systems among the Top500 systems worldwide as well as the specific situation in Europe. Based on the June 2013 Top500 list 269 of the 500 (i.e. 53.8%) systems are industrial. i.e. 53.8% of the systems belong to industry. The following figure illustrates the percentage of systems from various segments in the latest top 500.

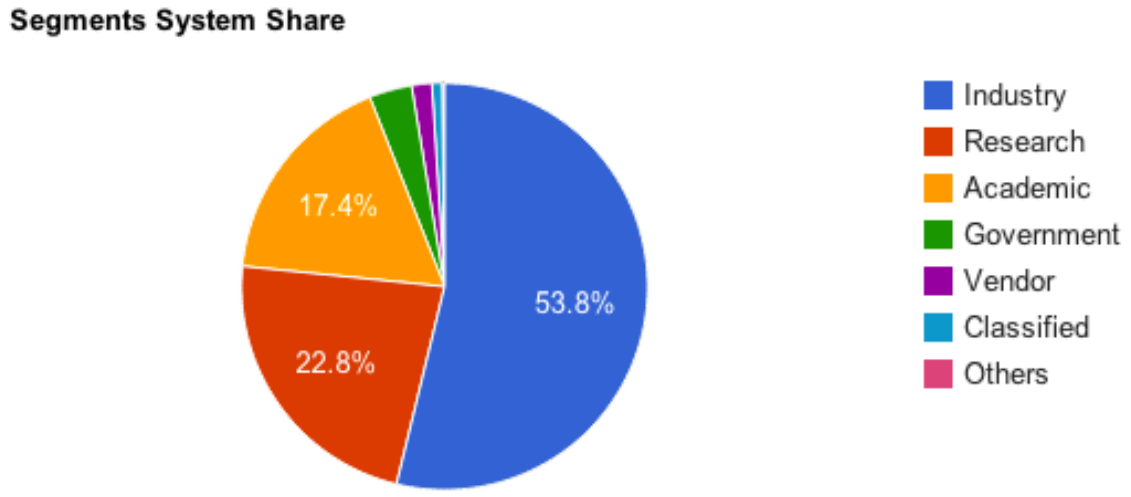

Figure 28- Segments System Share

However, the performance of those systems represents only 19.3 % of the total performance of all systems in the Top500 list as shown in the following figure.

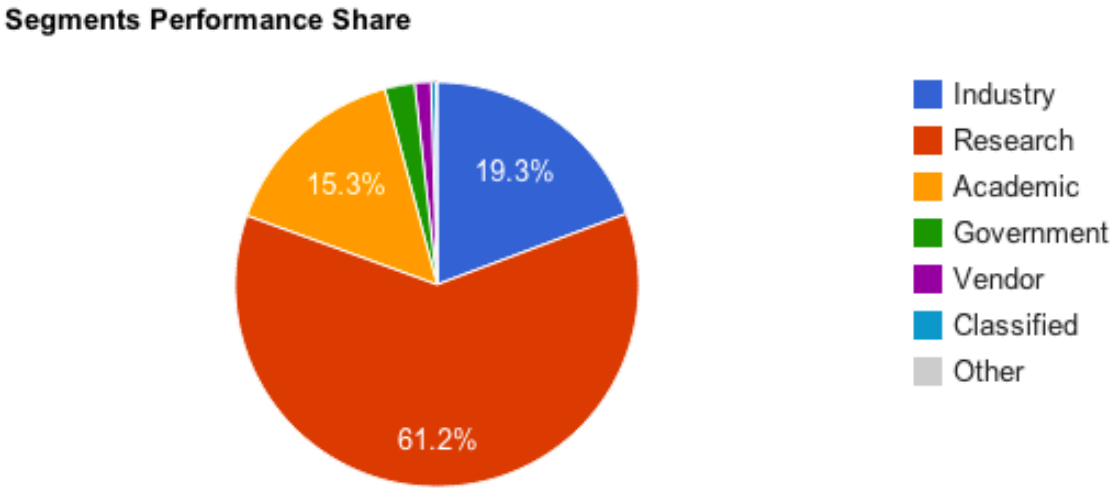

Figure 29 - Segments Performance Share

Moreover, in the Top 100 positions only 11 (11%) systems belong to industry and if we go even higher in the ranking, we can observer that in the Top 35 which contains all the Petaflop systems there are 2 systems that belong to industry.

It is worth mentioning that these 2 top systems both belong to European companies. More details about those will be provided later in this section.

Some indicative usage of all those top industrial systems are given below: Petroleum companies, Automotive, IT providers, Energy and Electricity, Financial and Banking, Airplane design, Electronics etc.

Further to the above statistics, 52 (19%) of the total industrial systems are located in Europe. However, as stated above, the two Petaflop industrial systems, delivered by US vendors, are in European sites. Those are:

- The Pangea - SGI ICE X, Xeon E5-2670 8C 2.600GHz, Infiniband FDR that is located in France and belongs to "Total Exploration Production". The system was inaugurated in March 2013 and will be used within the Seismic Imagery and Interpretation department of Total's Centre for hydrocarbon research. It will be used as a tool to assist decision-making in the exploration of complex geological areas and to increase the efficiency of hydrocarbon production in compliance with the safety standards and with respect for the environment.
- The second system is Hermit, a Cray XE6, Opteron 6276 16C 2.30 GHz, with Cray Gemini interconnect, that is located Germany and is owned by HWW and the university of Stuttgart. HERMIT is used by various industrial partners and is also one of the PRACE Tier-0 systems.

Although not a Petascale system, CEA CCRT "airain" supercomputer (Top500 Nr 192 in June 2013, currently a 200 Tflop/s bullx cluster) is worth mentioning because it is configured and operated for usage by a set of CEA industrial partners. Since 2003 CEA has proposed an original business model to industrials who can share the investments for a supercomputer, getting shares of the cycles and related HPC full services proportional to their investment. More than 10 industrial partners have now joined CCRT and use it for regular HPC production (in areas such as energy, aeronautics, automotive industry, electronics, cosmetics…). There are plans for upgrading airain in 2013 – and more generally following the demand of existing or new CEA partners.

4.3 Chapter Summary

The market watch of petascale systems conducted by PRACE since 2010 is now reaching the expected mid-way point in the "petascale era". The data is providing an interesting view of how this is shaping the high-end HPC market. This, together with the analysis of key players and their roadmaps provided by the business analysis, helps paint a picture of the current and near-future situation of petascale supercomputing hardware. It is now clear that the use of accelerators for reaching these high peak performances has carved a niche in the petascale market, yet its negligible growth has left it paired with traditional non-accelerated systems. The other hardware technique used for petascale, lightweight cores, is also seeing a spectacular growth, with IBM Blue Gene/Q and SPARC-based systems almost as popular, and with prospects of new ARM-based and other low-energy processors entering the high-end HPC market soon.

Vendors are aligning themselves with this reality, and are pushing their own versions of these three architecture, with almost all key players pursuing at least 2 of the methods, and many all three: Intel has already made its entry into the accelerated market with Xeon Phi, and the promise of totally independent units in Knights Landing could mean a complete Xeon Phi system (as lightweight processor, not as accelerator); AMD bought ATI to enter the accelerator market in 2006, and has now recently licensed ARM64 technology to dive into the lightweight-processor server market as well with Seattle. Could this be the definitive battle between x86 and ARM? Many other companies are devising strategies of their own, trying to keep all paths open while defending their interests at the same time.

5 Security in HPC centres

This chapter gives an update of the work done under security issues in the first half of 2013. The presented state of the art is covering topics which are important from an owner of HPC centre point of view. However it's not complete but is a good starting point to provide periodc security audits in each data centre. We refer to the white paper published on PRACE web site, stemming from these efforts, for further developments and references: this section is only a mere summary of the main outcomes of this effort.

5.1 The state of art brief summary

There are many security-related technologies used in HPC centres. Those of the greatest importance among others are the following:

- Network firewalls
- Antivirus software
- Local Intrusion Detection/Intrusion Prevention Systems
- Distributed Denial of Service protection
- Honeypots
- Data Loss Prevention / Data Leakage Prevention software
- Network segmentations – Demilitarized Zone, Virtual LANs
- Authentication
- Incident response procedure.

An **electronic survey was** circulated within PRACE consortium to explore these topics.

Local network firewalls

Guarding the entire network belongs to the tasks of the network firewall and therefore it is usually placed as close to the external network as possible. That prevents unnecessary network flow from occupying LAN devices' resources and therefore enhances performance. Instead of having just one firewall guarding the network, it is a common practice to use two or more firewall devices cooperating with each other. Redundancy may be additionally used together with load balancing.

Antivirus software

A software combating malware is popularly called anti-virus software, despite the fact that it detects and neutralizes all kinds of malicious software. Usually, it detects the malevolent software basing on signatures, i.e. known patterns of data within the executable code. It is therefore crucial to keep the signatures database updated.

Local Intrusion Detection/Intrusion Prevention Systems

Two main categories of Intrusion Detection Systems are:

- Host-based intrusion detection system (HIDS) – an application installed on a specified machine. Its main goal is to monitor certain operating system components as well as applications and network interfaces in order to discover suspicious activity that may be a sign of a break-in attempt.
- Network-based intrusion detection system (NIDS) – monitors network traffic and attempts to discover known attack patterns (signature-based approach) or unusual network activity (anomaly detection approach).

Intrusion Detection Systems with the active attack prevention mechanism and functionalities (blocking certain ports, resetting suspicious connection, etc.) are called Intrusion Prevention Systems.

Distributed Denial of Service protection

A specific group of network attacks that are worth mentioning are Denial of Service (DoS) attacks. Their concept bases on rejecting a legitimate user access to a certain service. A Distributed Denial of Service (DDoS) attack is a DoS attack that is additionally carried out from numerous different locations. It introduces two main factors: the attack volume is much higher and it is much more difficult (or even impossible) to define the list of attacking source IP addresses.

Honeypots

Sometimes knowing only that the attack happened is not enough, one wants to know how exactly it proceeded. The technology, which helps to achieve this, is a honeypot. It is a trap set to detect, deflect, or in some manner counteract attempts of unauthorized use of the system. It also allows gathering information about it for further analysis. Generally it consists of a host that appears to be part of a network, but is actually isolated and monitored. To attract the attackers it should also seem to contain valuable information or a resource.

Data Loss Prevention / Data Leakage Prevention software

DLP (Data Loss Prevention / Data Leakage Prevention) is the common name for a mechanism designed to control data transfers from a protected system to the external (public) area. It is especially dedicated to detect and, in some cases, prevent potential data leakage. One of the most important steps of setting up a DLP system is a process of defining sensitive data patterns. The other important step is the placement of the DLP system in the network structure. It should be placed in a point where a risk of sensitive data transfer from the inside of the organization to the public network exists.

Network segmentations – Demilitarized Zone, Virtual LANs

Proper network segmentation must not be omitted as well. It is recommended to place public services in a separate network segment called Demilitarized Zone (DMZ), having no access to the users' internal LAN. It can be achieved by either physical separation or using VLANs (Virtual LANs), which enable the administrators to divide the physical network into logical sub-networks. Except the easier network segmentation VLANs help, for example, to separate the user traffic from the administration traffic, which obviously increases the overall security level.

Authentication

Authentication basing on a username and a static password is, naturally, the most popular form of authenticating users. It is not, however, the most secure one. One of the relatively recent technologies that have been getting more attention lately is a one-time password method (OTP). OTP is a password that is valid for only one login session, which helps avoid some of the shortcoming of static passwords such as vulnerability to replay attacks, i.e. even if the attacker obtains the password, it is not valid and cannot be used anymore.

A better known technique is a use of asymmetric cryptography. A user needs a public and a private key for a successful authentication.

Incident response procedure

It describes a set of procedures describing the actions required after the occurrence of an event. Although this process does not require any particular technical expertise, it does require a lot of thoughts. Senior managers should carefully take into the consideration an incident response procedure after receiving a briefing based on the vulnerability assessment.

5.2 Chapter Summary – White Paper recommendations

The aforementioned PRACE survey fed a White Paper on HPC Centre Security. The process was started in the last stage of PRACE-1IP and continued in WP5 of PRACE-2IP. A released version of the White Paper is publicly available on the PRACE web site:

http://www.prace-ri.eu/IMG/pdf/wp79.pdf.

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

1. Perform security audits periodically
2. Perform formal audits of the organization
3. Introduce network security mechanisms (DMZ, VLANs, network firewalls, High Availability of firewalls, DLP systems, IDS and IPS systems)
4. Implement host security (host based firewall, antivirus software, DLP, host based IDS/IPS, policy for remote managements connections).

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.

Another important point is that security is a process, not a product. Security audits have to be performed regularly by, again, people not directly involved in maintaining of the infrastructure. Not all sites do perform configuration review and penetration tests on a regular basis; improving this situation is highly desirable.

The High-Performance Computing Centres present different levels of security. Some of them pay more attention to it while others seem to be more focused on performance. Although security always comes with costs, both, material and in the use of resources.The performance of a compromised system may tend to zero. Therefore important is to keep the security on the best possible level.

6 Exascalability: some trends and positions

Clearly, exascaling cannot be a goal in itself. Exascaling is driven by scientific and industrial questions that can only be solved by pushing the computational possibilities.

The Scientific Case published by PRACE last year (http://www.prace-ri.eu/PRACE-The-Scientific-Case-for-HPC) gives a multi-domain vision of such needs.

More recently in 2013 the Human Brain Project (cf. http://www.humanbrainproject.eu/) has also been selected as a European Future and Emerging Technologies Flagship project. Extrapolations from simulating a mouse brain using current state-of-the art supercomputers show that simulating a human brain requires among others exascale-computing resources that are envisioned for the 2020 timeframe.

Understanding the human brain is one of the greatest challenges facing 21st century science. If we can rise to the challenge, we can gain fundamental insights into what it means to be human, develop new treatments for brain diseases and build revolutionary new Information and Communications Technologies. Researching these challenges in neuroscience, medicine and computing, requires long-term investments.

The other selected FET Flagship project GRAPHENE (http://www.graphene-flagship.eu) will also probably be a strong driver for HPC and computational material sciences and nanotechnologies requirements.

Disclaimer:

This short chapter is a tentative effort to start further extrapolation of our watch and projection efforts, beyond the observation of current petascale consolidation and short-term trends. It relies on information and material that was mainly collected during or around ISC13 and about current EC FP7 projects. At ISC13 some vendors and technology providers were more visible or more active than others. It must also be said that more information was available this time from other providers, but under NDA with different WP5 members, and thus not suitable for comment in this deliverable.

We will continue and report on these topics in future Work Package 5 activities, building a more complete and balanced picture, with other vendors and providers input publicly available and probably new projects in Europe. This chapter must thus be seen as a current starting point.

6.1 Vision for Co-design and Fabric Integration

The vision of exascalability was presented by Alan Gara, the former Blue Gene (IBM) architect. It is a very good overview of several challenges which have to be done to reach a future system we can call as an exascale one [37].

Alan Gara says it is true to anticipate that technologies will play an important role in defining systems of the future and correspondingly the things users will need to do to extract performance. This is not as new of a direction as it might sound. Users have been adapting to the realities imposed by system architectures for a long time. The most obvious example is our inability to continue to increase frequency has resulted in users needing to exploit much larger degrees of concurrency.

There are a number of branch points in the future that technology will drive. There will be things that will dictate if we go one way or another, and none of us can predict today which way it will go. However, there are some things that we do anticipate and we know will be

there as part of all possible directions for exascale. It's really more a question of degree. Some technologies can sort of make the day – and make the switch easier.

The users should really be focusing on threading their applications to try to enable them, from a system architecture perspective, to exploit as much performance as possible from a finite amount of memory effectively – or a finite state of their problem. And the reason that's important is that memory itself is such a big swinger in the whole picture. Right now at the current systems the amount of silicon dedicated to memory is actually quite high and more silicon is involved in memory than there is in the processor. In addition it is also not scaling as fast as the performance of the computing is scaled.

Achieving Exascale will be an amazing accomplishment, which is likely to initially be focused on solving important highly scalable problems. The biggest challenge to reach Exascale is to do this in a manner that enables accessible performance, reasonable total system power, high reliability, and reasonable cost. And to achieve this in a reasonable timescale. It is known how to do each of these in isolation but doing all simultaneously represents the real challenge.

Memory comes into the Exascale challenge in a number of ways. The most important dimension is energy efficiency. This is more a memory microarchitecture innovation as opposed to a fundamentally new physical device. The exascale needs to dramatically reduce the energy that is needed to access memory. Of course there is also the possibility that new device technology could also help energy efficiency. Right now though most of the energy associated with memory is not attributable to the actual physical memory cell.

New memory technologies are extremely important for the future. The scaling of the physical DRAM device is getting much more difficult going forward. The memory density improves at a much slower rate than the increase of compute performance. This has put extreme stress on users and without new memory technologies this skewing will continue.

Threshold Voltage really offers an opportunity to get significant improvements in energy efficiency at the transistor level. It plays a very important role when looking at near threshold carefully. Near Threshold Voltage comes at a pretty significant decrease in the performance of those devices. The amount of silicon area per device and the performance of that device both go down. The reality is the energy efficiency at the system level.

There is a need to take a broad system view in assessing these technologies. It is not just the question of how efficient a single transistor is but really how efficient is a system for real applications that is built out of such transistors.

In other words, important is to transition the thinking from energy efficiency at the transistor level – to energy efficiency at the system level.

When exploring the question of 'Does Near Threshold Voltage show promising results for exascale', getting to an answer is much more complex than a simple yes or no as it makes assumptions as to what user applications will look like in 5 to 10 years. If the only requirement is to build a system that could achieve 1 EFlop/s for a simple code that it could be achievable by the end of the decade. But it does not make much sense to build a machine that is not highly usable so the degree to push in directions like near threshold voltage is tempered by this. The long term answer to **this will be to operate in many different domains, i.e to operate at very low voltages when the application can exploit extreme levels of parallelism and to operate in a mode which is optimal for algorithms that have far less parallelism**.

The industry has a long history of absorbing things that were one day considered accelerators into part of the baseline architecture. One example is the floating point units. These used to be add-on accelerator devices much like GPUs are today. On the other hand, accelerators like GPUs have been fairly difficult for the community to use. They have been explored in HPC for more than a decade and there remain very few production codes which have shown better performance. Some of this is due to them not being integrated more closely into the processor. Valued features/concepts will be integrated into a CPU where it makes sense. Integrating accelerators is a viable direction that is explored but they need to have enough of an application reach to justify the silicon area.

Fabric integration is one of the natural next steps to go. Co-design is fundamental being able to build systems that are usable, cost effective and power efficient.

6.2 Hardware development and basic R&D

The hardware requirements were collected mainly at the ISC 2013 conference and exhibition in Leipzig, the HPC workshop in Lugano. There were also direct meetings and discussions with HPC vendors, like Bull, Cray, HP, IBM, Intel, SGI but also data infrastructure vendors like Panasas, DDN and technology providers like Samsung.

The current Top500 is x86 dominated: 80% Intel, 10% AMD. Just looking at the new entries, Intel was used in 174 out of 177 systems. IBM's Blue Gene/Q is still the most popular system in the TOP10 with four entries. Including Tianhe-2, the current TOP1, there are two Intel Xeon Phi accelerated systems. Furthermore there are two NVIDIA accelerated systems in the TOP10. The top systems reach some 2 Gflop/W with a total energy budget of some 20 MW. The max energy consumption of an exascale system was set on the threshold of 20 MW (HPC workshop in Lugano) and it seems like this is a value presented also by several hardware vendors at ISC 2013.

All of the mentioned above vendors are working on issues which will allow us to deploy and use efficient, reliable and highly efficient exascale systems. Because the exascale perspective is much beyond the near future (the usually available roadmaps describe 2-3 years in advance) and covers 2018-2020, most of the information are under NDA and not available at all. This is the reason we present only information from two vendors, which are publicly available.

6.2.1 Intel Road to Exascale

At ISC'13, Rajeeb Hazra of Intel gave a presentation "Driving Industrial Innovation on the Road to Exascale". Although Intel is not producing or integrating complete HPC systems, around 80% (403 out of 500) of the current (June 2013) Top500 entries are systems powered by Intel components.

Moore's law is alive and well. Coming from 32 nm (2009) via 22 nm to 14 nm (2013) Intel expects to move to 10 nm (2015+), 7 nm, and 5 nm. Going forward with scaling, new materials and device structures are needed. The circuit design and micro-architecture innovations will focus more on power efficiency.

With Xeon Phi (early 2013), Intel re-introduced co-processors as accelerators. The product has not only been introduced but has also been demonstrated at scale in the current Top500 #1, the Tianhe-2. The next generation Xeon Phi, code name Knights Landing, according to the current plans of Intel, will use 14 nm technology, will be able to run either as standalone CPU or as PCIe co-processor and will have integrated on-package memory.

Intel is not only active in the area of processors but is also active on interconnects. After acquiring QLogic InfiniBand, Intel also acquired Cray's Gemini/Aries technology and is working on the next generation of the Intel True Scale interconnect.

Looking at today's technology, there is already an integration of math co-processors, graphics, I/O controllers, memory controllers and on-package memory. Possibilities for tomorrow are further integration with: fabrics, storage, and switches. This will be accomplished using 3D chip stacking, system integration and silicon photonics and requires hardened circuits and architecture with respect to resiliency.

The one exception to scaling with Moore's law is DRAM memory. Here, the trend is to move to smaller physical memory sizes and to cope with lesser memory per computing element via threading. Furthermore, investment and innovation in new high-density memory technology is needed (which is a separate topic in this report).

From a software point of view, there are two steps: improve thread scalability performance, and the need for new memory architectures and storage models for the new high-density memory technologies.

According to Intel there are two design options for exascale computing:

1. Having more cores and/or a variety of cores, having larger cache but external memory. This would result in a large die with more than 10 billion transistors.
2. Using fewer cores with on-package memory.

6.2.2 Hardware components improvement by HP

The top priorities defined by HP Labs for future exascale systems are following [35, 36]:

- Improve Performance/TCO by 10X–
- Efficiency:
  - **Interconnects using photons**
    - 5x (short term: 5years) optical links between nodes
    - 10x (long term) with nanophotonics (+10x bandwidth)
  - Nodes with 256 cores : 10TFlop/s per200Watts
  - Memory hierarchy extended with **memristors**
- Manage: 1 operator for100K nodes
- Autodetect and autorepair failures:
  - Check-point Restart integrated and transparent

Four research axes as priorities:

- Optical interconnects: Scalability up to 1M nodes
- Basic blocks for compute: Corona project
- System software: 1 operator for100K nodes
- Programmability: Reliability, efficiency

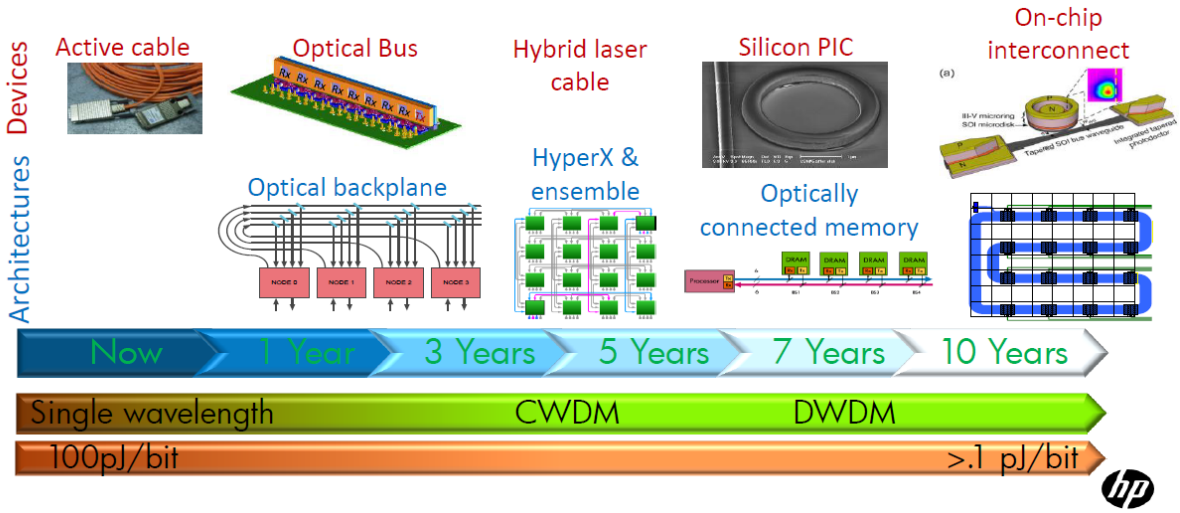

Figure 30 - Photonics technology in 5-10 years perspective

Figure 31 - The order of introducing memristor technology

The memristor is a basic element of future memory revolution:

- Scaling down to less than 10 nm width per cell (~32 GB/cm2/layer by 2018)
- Scaling up to multiple (more than 8) layers on chip (~0,25 TB/cm2/chip by 2018
- Truly nonvolatile for many years
- Random Access
- Fast cell write and erase (~ns)
- Low energy cell write and erase (picoJ)
- Good to excellent endurance

The goal is to reach a production technology of non-volatile (NV) memory and storage in 7-8 years (50G$/y).

6.2.3 Energy 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. The commonly agreed goal of an exascale supercomputer is within a power envelope of 20 megawatts (MW). On the other hand, extrapolating Beacon, the previous No. 1 supercomputer on the Green500, exascale would result in a 408 MW machine. This gives us an impression how far is the current technology from the expected goal.

Following new innovative approaches in energy efficiency of Eurora, the new No. 1 supercomputer on the Green500, this exascale power envelope comes down to 312 MW, a sizeable 24% drop in power consumption for an exascale machine. Nevertheless, the electricity bills for such a system would still be more than 200 million Euros per year.

Power efficiency challenges by data movement between chips

In the context of energy efficiency at exascale level an issue of growing concern is towards data movement between chips. Off-chip communication is a primary scalability limiter for scientific computing as well as in addition to limiting performance off-chip communication also consumes considerable power. With current technologies, the power required to move a bit off-chip is nearly two orders of magnitude larger than that required moving a bit on- chip.

Silicon photonics is a maturing technology that has the potential to improve performance and reduce power consumption. Photons are currently much more power efficient than electrons for moderate and long distance communication and are widely used when communicating over distances larger than several meters, e.g. rack-to- rack communications.

The advantages of optics include higher wiring density and a power cost that does not grow with distance. But the power required to convert between electrons and photons has limited the advantages of optical communication for shorter links (e.g. within a rack, board or chip). Approaches to address these limitations are active areas of research in both components and systems, and promising prototypes have been demonstrated in laboratory settings. Progress in accelerating the maturation of optical networking would be beneficial to multiple computing communities.

6.2.4 Memory Reliability

Resiliency will be one of the toughest challenges in future exascale systems. Memory errors contribute more than 40% of the total hardware-related failures and are projected to increase in future exascale systems.

Error correction codes (ECCs) and checkpointing are two effective ways to protect a system from memory induced failures. By detecting and correcting memory errors on the fly, ECCs can significantly increase the reliability of the memory subsystem and thus increase the overall system resiliency. While ECCs protect systems from memory errors before they cause system failures, checkpointing recovers systems from failures not averted by the ECCs. Although warehouse data centers may allow individual nodes to be off-line when machines fail, exascale supercomputing leverages checkpointing to handle system failures since nodesare usually working together to solve large scale problems. In the presence of a failure, computation will restart from the last checkpoint, which wastes previous work on all the unaffected nodes.

6.3 EU project examples

There are currently 3 projects launched by EC under Framework Programme 7, two years ago, to start prototyping exascale concepts.

6.3.1 DEEP

The DEEP – Dynamical Exascale Entry Platform – project (cf. http://www.deep-project.eu/) is an Exascale project funded by the European 7th FP. The project DEEP will develop a novel, Exascale-enabling supercomputing platform along with the optimisation of a set of grand-challenge applications highly relevant for Europe's science, industry and society. The DEEP

System will realise a Cluster Booster Architecture that will serve as proof-of-concept for a next-generation 100 Pflop/s production system.

The current DEEP system comprises a 128 node Eurotech Aurora Cluster and a 256 node Booster. The Booster in itself is a cluster of accelerators. This differs from the more conventional accelerator(s) per node approach. A Booster node contains: two Intel Xeon Phi processors and two EXTOLL NICs for building a 3D torus interconnect. (EXTOLL is a Heidelberg University spin-off, originally using FPGAs.) The complete system is Direct Liquid Cooled with an up to 50 °C inlet temperature.

The DEEP software stack contains ParTec's ParaStation MPI and exploits a task-based programming paradigm based on OmpSs (originating from BSC).

The current (June 2013) Green500 entries #1 and #2 are both Eurotech systems:

1. Eurora at CINECA reaching 3.21 Gflop/s/W
2. Aurora Tigon: 3.18 Gflop/s/W

6.3.2 Mont-Blanc

The Mont-Blanc project (cf. http://www.montblanc-project.eu/) is an exascale project funded by the European 7th FP. It has set itself the following objective: to design a new type of computer architecture capable of setting future global HPC standards that will deliver Exascale performance while using 15 to 30 time less energy than the current trends would lead to.

Mont-Blanc tries to leverage commodity and power-efficient technology. Current commodity is cell phones (specifically smartphones) and tablets. The idea is that as microprocessors killed the vector supercomputers (just look at the current Intel share in the Top500), history may be about to repeat itself: mobile processors are not faster but are significantly cheaper.

The current Mont-Blanc prototype uses Samsung Exynos 5 dual compute cards (having two Cortex-A15 SoCs running at 1.7 GHz and a Mali T604 GPU). For near future systems there are several interesting upcoming SoCs: Exynos 5 Octa, Tegra 4, Snapdragon 800, … The prototype uses a bullx carrier blade having 15 compute nodes with an integrated GbE switch.

The current prototype reaches 2.4 Gflop/W.

6.3.3 CRESTA

The CRESTA project (Collaborative Research into Exascale Systemware, Tools & Applications, http://cresta-project.eu/) is a 7 FP EU project with the aim to provide exascale requirements from the end user point of view and new tools and systemware.

Having demonstrated a small number of scientific applications running at the petascale, the nature of the HPC community, particularly the hardware community, is to look to the next challenge. In this case the challenge is to move from 10^{15} Flop/s to the next milestone of 10^{18} flop/s – an exaflop. Hence the exascale challenge that has been articulated in detail at the global level by the *International Exascale Software Project* and in Europe by the *European Exascale Software Initiative*. Many of the partners in CRESTA are leading members of one or both of these initiatives.

In tackling the delivery of an exaflop/s formidable challenges exist not just in scale, such systems could have over a million cores, but also in reliability, programmability, power consumption and usability (to name a few).

The timescale for demonstrating the world's first exascale system is estimated to be 2018. From a hardware point of view we can speculate that such systems will consist of:

- Large numbers of low-power, many-core microprocessors (possibly millions of cores)
- Numerical accelerators with direct access to the same memory as the microprocessors (almost certainly based on evolved GPGPU designs)
- High-bandwidth, low-latency novel topology networks (almost certainly custom-designed)
- Faster, larger, lower-powered memory modules (perhaps with evolved memory access interfaces)

Only a small number of companies will be able to build such systems. However, it is crucial to note that hardware is not the only exascale computing challenge, but also software and applications.

6.4 Chapter Summary

The exascale trends can be observed on many levels:

- Big Computing
  - Millions of cpus and cores
- Big Data
  - Data tsunami
- Networking
  - Faster interconnects.

The chapter was concentrating on computing and interconnects, without the analysis of big data issues. However, the data management is one of the key topics.

On the other hand exascale is not a one criteria development. We have to utilize the computing and data infrastructure based on applications, which will be available in 5-7 years. Therefore it is very important to find the applications, their specific requirements and see whether the architecture can be called as future exascale (full utilization) or if we are talking only about one application solution (adopted and fast but only for a dedicated application).

Therefore it is very important to create centres of competence (CoC) and joined EU projects (academia + industry). There are several examples of EU international projects: Mont-Blanc, CRESTA or DEEP. We also see several CoC:

- IBM Exascale Innovation Centre, Jülich
- Intel Exascale labs
- MAQAO, Scalasca, Paraver
- Bull and Intel (Center for Excellence in Parallel Programming)

Throughout this chapter but also in previous chapters, we observe a couple of exascale trends.

- There is an ongoing trend of most vendors towards hybridization of CPU and accelerators. This trend is in both directions: CPU vendors adding accelerator cores and GPU vendors adding CPU like cores.
- There is a trend to even higher levels of integration.
- Exascale cannot be reached without using more cores/threads per available memory.
- Similar to the slow but steady takeover of the Top500 by PC commodity x86 CPUs, we might as well expect a new takeover by the new cell phone commodity SoCs.

- Reaching exascale at affordable electrical power costs. The generally accepted DARPA exascale goal is using a maximum energy budget of 20 MW.
- There is a trend towards using integrated programming models. The problem here is that we still have no indication which programming model will prevail. This resembles the early days of message passing where every vendor had its own message-passing library. After the introduction of MPI in the early 1990's it still took years before the dust settled.

7 PRACE and the European HPC Ecosystem in a Global Context

In February 2012 The European Commission published a communication that underlines the strategic nature of HPC [42] . 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.

This communication and its perspectives were at the agenda of a Council of Competitiveness meeting, May 30th, 2013 [27]. The conclusions are a clear recognition of the need for an EU-level policy in HPC addressing the entire HPC ecosystem. PRACE and ETP4HPC are recognized as key players of this ecosystem, resp. at the infrastructure level and at the technology supply level. A world-class and sustainable HPC infrastructure is indeed considered crucial, as well as HPC industrial supply for development of exascale computing and excellence in HPC software, methodology and applications, for HPC use by Science and by industry, including SMEs. For this latter pillar of usages and applications, European-wide Centres of Excellence and networks in HPC applications addressing key societal, scientific and industrial challenges in areas that are strategic for Europe are foreseen; as well as more national or regional HPC Competence Centres to support the transfer of relevant expertise from supercomputing centres to industry – including to SMEs.

At the eve of Horizon 2020, discussions on the precise mechanisms to implement such a strategy are still on going.

7.1 PRACE

PRACE is regularly delivering cycles – every 6 months through the so-called Regular Calls - on the 6 petascale, Tier-0 systems of its Hosting Members, accounting for an aggregated peak performance of circa 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

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 all these 6 systems were registered in June 2013 Top500 list [28], all in the Top50, resp. at ranks 7, 9, 12, 15, 29 and 32 (2 of these systems are in the Top10 and 4 in the Top20).

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 or Chinese projects – Tianhe-2, Titan@ORNL, BlueWaters@NCSA but also STAMPEDE@TACC, i.e. GPGPU or manycore/MIC fuelled systems - that PRACE does not have 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 or Intel Xeon Phi (MIC) clusters are currently available within PRACE systems, and one Tier-0 system only hosts GPUs, made available to Regular Calls every six month. The most recent hybrid configuration made available within PRACE is CINECA's EURORA Xeon Phi partition of circa 100 Tflop/s (by EUROTECH), open to PRACE internal usage mid-2013 for code petascaling (PRACE 1IP project activity). Let us note also that EURORA's GPU partition – using NVIDIA K20 beside Intel Sandy Bridge - is ranking Top1 in June 2013 Green500 [29], [30], a clear evidence of Europe's dynamism in the area of energy-efficient HPC.

Regarding usages, since mid-2010 PRACE has been maintaining a steady growth from 363 to the order of 1200 to 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. Not strictly comparable with PRACE in term of frequency and duration of calls and allocations, we can however get an idea of global number through what was posted for INCITE at:

https://proposals.doeleadershipcomputing.org/allocations/calls/incite2014

"INCITE is currently soliciting proposals of research for awards of time on the 27-petaflop/s Cray XK7 "Titan" and the 10-petaflop/s IBM Blue Gene/Q "Mira" beginning Calendar Year (CY) 2014. More than five billion core-hours will be allocated for CY 2014. Average awards per project for CY 2014 are expected to be on the order of 50 million core-hours for Titan and 100 million core-hours for Mira, but could be as much as several hundred million core hours. Proposals may be for up to three years."

NB: this call is now closed - INCITE proposals are accepted between mid-April and the end of June. We can thus roughly estimate INCITE to be 2 times bigger than PRACE if we compare on one running year (1 INCITE call vs. 2 PRACE calls).

For instance PRACE Regular Call 7 offered the following amounts of millions of core*hours on PRACE tier-0 systems, in July 2013 – for projects starting October 2013:

- FERMI = 480
- JUQUEEN = 100
- MARE NOSTRUM = 110
- CURIE Fat Nodes = 28
- CURIE Thin Nodes = 201
- CURIE Hybrid (GPU) = 0.5
- HERMIT = 120
- SUPERMUC = 220

Regular 8 allocations – call open September-October 2013, allocation decisions February 2014, start of projects March 2014 – should be in the order of:

- •FERMI = 480
- •JUQUEEN = 100
- •MARE NOSTRUM = 110
- •CURIE Fat Nodes = 28
- •CURIE Thin Nodes = 201

- • CURIE Hybrid (GPU) = 0.5
- • HERMIT = 120
- • SUPERMUC = 220

Beside the Tier-0 calls PRACE Implementation Phase Projects are announcing also regular Tier-1 calls for European scientific and industry communities (DECI programme).

PRACE is already 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.

7.2 ETP4HPC

ETP4HPC has been formally established by the European Commission as one of the recognised European Technology Platforms (ETPs). ETP4HPC is now included in the list annexed to the strategy for European Technology Platforms - ETP 2020 [31]. This makes ETP4HPC a distinguished voice for the definition of European HPC priorities and related R&D&I programmes.

ETP4HPC had previously released its Strategic Research Agenda in April 2013 [32] and presented it during ISC13 in Leipzig in June.

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 terms 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.

This research programme provides contents that should be turned into research topics for the first Work Programmes defined by the European Commission for 2014-2105 first period of Horizon 2020.

PRACE and ETP4HPC are complementary and have established a constructive dialogue so that their responsibilities are clearly divided reflecting their respective domains and expertise. Both organisations are discussing, as well as with EC DG-CONNECT, how to co-operate more closely to strengthen Europe’s place on the global HPC stage.

8 Conclusion and summary

The deliverable D5.2 is a result of work package WP5 work within PRACE-2IP project and summarizes the technologies and analysis of features that are the most important from the Tier-0 and Tier-1 sites point of view:

- Features worth to be taken into account while preparing a procurement specification
- Important issues for building a Tier-0 or Tier-1 data centre
- Architectures and technologies for Petascale systems
- Technologies that will lead to Exascale system in 2018-2020.

Some of the research is a continuation of topics from deliverable D5.1, i.e.

- Assessment of petascale systems
- Hardware requirements and trends
- PRACE and the European HPC Ecosystem in a Global Context
- Energy efficiency in HPC
- Cooling systems and its efficiency
- Power Measurement Methodology.

In addition to the updated topics, the D5.2 includes also completely new subjects like exascalability and infrastructure monitoring/management platforms and goes into much more details for infrastructure issues.

The ISC 2013 exhibition, conference and direct discussions with companies resulted in gathering information from following topics:

- HPC hardware requirements
- Petascacale architectures
- Assessment of petascale systems
- New CPUs
- New HPC architectures
- Exascalibility
- Energy efficient systems
- Green IT
- Solutions for data centres
- Technologies for data centres
- Cooling
- Water cooling and heat re-use
- Energy and heat monitoring systems
- Management software / operations
- Grand challenges.

Another notable result of WP5 was the publication of the final report on the security (White Paper on Security in HPC Centres), which is publicly available on the PRACE web site together with previous white papers on HPC centres infrastructures design, procurement and operations. As rightly noted in the white paper, the security is not a product or service, but a process that is aging and as time passes becomes outdated. It means that the process to maintain an adequate level of security should be updated periodically. We think the most important elements of this process are presented in the form of the features which should be reflected in the security policy and implemented by security audits, deployment of network and host based firewalls, antivirus software, IDS/IPS systems, DLPs, authentication and incident response procedures.

An undeniable success of the work packages WP5 and WP9 is the continuation of joint workshops on technologies for HPC centers, organized by the CEA (France), LRZ (Germany) and CSCS (Switzerland). In April 2013, the 4th edition of this event took place in Lugano (Switzerland).

Exascalability technology is certainly worth examining in the context of data centers development, especially building infrastructures. Here the technology development of computer systems is still a far future, because we know product plans are presented for two, up to three years, while the exascalability in the development of data centers must be anticipated with some present issues already. The first exascale machine will be released in the 2018-2020 time frame. This means that changes in the infrastructure of data centers can be very far-reaching, and should be already planned today. However, newly built data centers need to have at this time to predict the development of technologies to take into account the possibility of the development of big systems, in particular Tier-0 ones. One factor is of course the size of the energy power needs, the possibility of energy recovery (Green IT), saving on cooling, thereby reducing the PUE, the type of cooling technology, which depends directly on the scale of integration, etc. Another trend on how exascalability should be emphasized and carefully considered is the big amounts of expectable raw data and final results or visualisation data. The same processing will not be possible without a concentration of effort on the combination of data and compute infrastructure.

The exascale trends can be oberved on many levels: Big Computing (milions of cpus and cores), Big Data (data tsunami), network and communication (faster interconnects).

It is very important for the exascale perspective to find the right benchmark applications, as we have to utilize the computing and data infrastructure based on applications which will be available and deployed in 5-7 years.

Today 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.

This variety of system architectures of PRACE Tier-0 systems (together with Tier-1, similar to Tier-0 ones but at a smaller scale) gives users a good chance of finding the right system for their today petascale applications.

This report collected elements that can be useful to PRACE on its way to future generation systems in the next few years, from multi-petascale to exascale, regarding options and orientations for HPC architectures, their hosting technical infrastructures, and related design issues.