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

D5.3
Updated Best Practices for HPC Procurement and Infrastructure

Final

Version:	1.0
Author(s):	Andreas Johansson, SNIC-LiU
Date:	22.08.2014

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: < **D5.3**>	
	Deliverable Nature: Report	
	Deliverable Level: PU	**Contractual Date of Delivery:** 31 / 08 / 2014
		Actual Date of Delivery: 31 / 08 / 2014
	EC Project Officer: Leonardo Flores Añover	

Document Control Sheet

Document	**Title: Updated Best Practices for HPC Procurement and Infrastructure**	
	ID: D5.3	
	Version: <1.0>	**Status:** ***Final***
	Available at: http://www.prace-project.eu	
	Software Tool: Microsoft Word 2007	
	File(s): D5.3.docx	
Authorship	**Written by:**	Andreas Johansson, SNIC-LiU
	Contributors:	Guillermo Aguirre de Cárcer, BSC François Robin, CEA Jean-Philippe Nominé, CEA Eric Boyer, CINES Bartosz Kryza, Cyfronet Łukasz Dutka, Cyfronet Marco Sbrighi, GINECA Ioannis Liabotis, GRNET Marianthi Polydorou, GRNET Josip Jakic, IPB Norbert Meyer, PSNC Radek Januszewski, PSNC Gert Svensson, SNIC-KTH Walter Lioen, SURFsara Michał Białoskórski, TASK Mścisław Nakonieczny, TASK
	Reviewed by:	Herbert Huber, LRZ Thomas Eickermann, FZJ / PRACE PMO
	Approved by:	MB/TB

Document Status Sheet

Version	Date	Status	Comments
0.5	01/August/2014	Draft	Internal WP review
0.6	05/August/2014	Draft	Added chapter 3, executive summary, introduction and conclusion.
0.7	08/August/2014	Draft	Addressed WP internal review comments
1.0	22/August/2014	Final version	Addressed review comments

Document Keywords

Keywords:	PRACE, HPC, Research Infrastructure, Petascale, Exascale, Data Centre, Cooling, Electricity, Monitoring, Big Data, Interconnects, Top500, Green500

Table of Contents

List of Figures

List of Tables

References and Applicable Documents

[1] PRACE-2IP Deliverable D5.1, http://www.prace-ri.eu/IMG/pdf/d5.1-2.pdf
[2] PRACE-2IP Deliverable D5.2, http://www.prace-ri.eu/IMG/pdf/d5.2-2.pdf
[3] ISC'14 web page, http://www.isc-events.com/isc14/
[4] Heroux, M., Dongarra, J., Toward a New Metric for Ranking High Performance Computing Systems, http://www.netlib.org/utk/people/JackDongarra/PAPERS/HPCG-Benchmark-utk.pdf
[5] Fujitsu FEFS presentation, http://www.fujitsu.com/downloads/TC/sc11/fefs-sc11.pdf
[6] http://www.storageclarity.com/GRAU_DataSpace_Data_Sheet.pdf
[7] http://www.stornext.com/deployments/new-features-in-stornext-5/
[8] http://hpccloud.tistory.com/attachment/cfile22.uf@16207948513EBBA106644D.pdf
[9] Scality web page, http://www.scality.com/
[10] GPFS/Elastic Storage, http://www.ibm.com/systems/platformcomputing/products/gpfs/
[11] PanFS web page, http://www.panasas.com/products/panfs
[12] Ahern, Sean, Shoshani Arie, Ma Kwan-Liu, Choudhary Alok, Critchlow Terence, Klasky Scott, Pascucci Valerio, Ahrens Jim,Bethel Wes E., Childs Hank, Huang Jian, Joy Ken, Koziol Quincey, Lofstead Gerald, Meredith Jeremy S., Moreland Kenneth,Ostrouchov George, Papka Michael, Vishwanath Venkatram, Wolf Matthew, Wright Nicholas, and Wu Kensheng, *Scientific Discovery at the Exascale*, a Report from the DOE ASCR 2011 Workshop on Exascale Data Management, Analysis, and Visualization, ASCR, 2011
[13] M. Atkinson et al. Data-Intensive Research Workshop Report. Technical Report, e-Science Institute, access:08.02.2012, 2010
[14] Dongarra, J., Beckman, P. H., Moore, T., Aerts, P., Aloisio, G., Andre, J.-C., Barkai, D., Berthou, J.-Y., Boku, T., Braunschweig, B., Cappello, F., Chapman, B. M., Chi, X., Choudhary, A. N., Dosanjh, S. S., Dunning, T. H., Fiore, S., Geist, A., Gropp, B., Harrison, R. J., Hereld, M., Heroux, M. A., Hoisie, A., Hotta, K., Jin, Z., Ishikawa, Y., Johnson, F., Kale, S., Kenway, R., Keyes, D. E., Kramer, B., Labarta, J., Lichnewsky, A., Lippert, T., Lucas, B., Maccabe, B., Matsuoka, S., Messina, P., Michielse, P., Mohr, B., Müller, M. S., Nagel, W. E., Nakashima, H., Papka, M. E., Reed, D. A., Sato, M., Seidel, E., Shalf, J., Skinner, D., Snir, M., Sterling, T. L., Stevens, R., Streitz, F., Sugar, B., Sumimoto, S., Tang, W., Taylor, J., Thakur, R., Trefethen, A. E., Valero, M.; van der Steen, A., Vetter, J. S., Williams, P., Wisniewski, R. W. & Yelick, K. A., 2011, "The International Exascale Software Project roadmap", IJHPCA 25 (1) , 3-60
[15] Steve Ashby, et. al., The Opportunities and Challenges of Exascale Computing, 2010
[16] Lustre project web page, http://wiki.lustre.org/

[17] Intel Federal LLC, EXTREME-SCALE COMPUTING RESEARCH AND DEVELOPMENT FAST FORWARD STORAGE AND I/O MILESTONE: 8.5 – Final Report, Available online: https://wiki.hpdd.intel.com/download/attachments/12127153/M8.5%20FF-Storage%20Final%20Report%20v3.pdf?version=1&modificationDate=1404771975924&api=v2
[18] Exascaler product website, http://www.ddn.com/products/lustre-file-system-exascaler
[19] Jülich Supercomputing Centre tackles the grand challenges of research, IBM Systems and Technology Case Study, July 2013, http://public.dhe.ibm.com/common/ssi/ecm/en/xsc03152usen/XSC03152USEN.PDF
[20] I. Raicu, I. T. Foster, and P. Beckman. Making a case for distributed file systems at Exascale. In Proceedings of the third international workshop on Large-scale system and application performance (LSAP '11), 2011, ACM, New York, NY, USA, 11-18.
[21] Salem El Sayed, Stephan Graf, Michael Hennecke, Dirk Pleiter, Georg Schwarz, Heikos Schick, Michael Stephan, Using GPFS to Manage NVRAM-Based Storage Cache, In Supercomputing, LNCS, Volume 7905, 2013, pp 435-446.
[22] Exascale10 workgroup home page, http://www.eiow.org/
[23] Andre Brinkmann, Toni Cortes, Hugo Falter, Julian Kunkel, Sai Narasimhamurthy, E10 – Exascale10, E10 Whitepaper, 26/05/2014
[24] Inifinite Memory Engine product website, http://www.ddn.com/products/infinite-memory-engine-ime
[25] J. He, J. Bennett, and A. Snavely. DASH-IO: an empirical study of flash-based IO for HPC. In Proceedings of the 2010 TeraGrid Conference (TG '10). ACM, New York, NY, USA, 2010.
[26] Augusto Burgueño Arjona, DG-CONNECT, http://www.etp4hpc.eu/wp-content/uploads/2014/06/ETP4HPC-ISC14-BOF-session.pdf

List of Acronyms and Abbreviations

AISBL	Association International Sans But Lucratif (legal form of the PRACE-RI)
AMD	Advanced Micro Devices
API	Application Programming Interface
ASHRAE	American Society of Heating, Refrigerating and Air-Conditioning Engineers
ASIC	Application-Specific Integrated Circuit
ATI	Array Technologies Incorporated (AMD)
AVX	Advanced Vector Extensions
AWE	Atomic Weapons Establishment
Big-endian	Big-endian is the ordering when the most significant byte in a value is stored first (at the lowest storage address)
BLAS	Basic Linear Algebra Subprograms
BSC	Barcelona Supercomputing Center (Spain)
CEA	Commissariat à l'Énergie Atomique 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)
CPU	Central Processing Unit
CSCS	The Swiss National Supercomputing Centre (represented in PRACE by ETHZ, Switzerland)
CUDA	Compute Unified Device Architecture (NVIDIA)

CUE	Carbon Usage Effectiveness
DARPA	Defense Advanced Research Projects Agency
DDN	DataDirect Networks
DDR	Double Data Rate
DDR4	Fourth generation DDR
DECI	Distributed European Computing Initiative
DEISA	Distributed European Infrastructure for Supercomputing Applications. EU project by leading national HPC centres.
DIMM	Dual Inline Memory Module
DKRZ	Deutsches Klimarechenzentrum
DMA	Direct Memory Access
DP	Double Precision, usually 64-bit floating point numbers
DRAM	Dynamic Random Access memory
DSP	Digital Signal Processor
EC	European Community
EDR	Enhanced Data Rate, 25 Gbit/s InfiniBand
EESI	European Exascale Software Initiative
EoI	Expression of Interest
EP	Efficient Performance, e.g., Nehalem-EP (Intel)
EPCC	Edinburg Parallel Computing Centre (represented in PRACE by EPSRC, United Kingdom)
EPFL	École polytechnique fédérale de Lausanne
EPSRC	The Engineering and Physical Sciences Research Council (United Kingdom)
ERE	Energy Reuse Effectiveness
ETHZ	Eidgenössische Technische Hochschule Zürich, ETH Zürich (Switzerland)
EX	Expandable, e.g., Nehalem-EX (Intel)
FC	Fibre Channel
FFT	Fast Fourier Transform
FP	Floating-Point
FPGA	Field Programmable Gate Array
FPU	Floating-Point Unit
FRAM	Ferroelectric Random Access Memory
FZJ	Forschungszentrum Jülich (Germany)
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)
GENCI	Grand Équipement 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
GNU	GNU's not Unix, a free OS
GPGPU	General Purpose GPU
GPI	Fraunhofer implementation of the Global Address Space Programming Interface
GPU	Graphic Processing Unit
HBA	Host Bus Adapter
HCA	Host Channel Adapter

HDD	Hard Disk Drive
HE	High Efficiency
HP	Hewlett-Packard
HPC	High Performance Computing; Computing at a high performance level at any given time; often used synonym with Supercomputing
HPL	High Performance LINPACK
HSM	Hierarchical Storage Management
HT	HyperTransport channel (AMD)
HTTP	HyperText Transfer Protocol
IB	InfiniBand
IBM	Formerly known as International Business Machines
ICE	(SGI)
IDRIS	Institut du Développement et des Ressources en Informatique Scientifique (represented in PRACE by GENCI, France)
IEEE	Institute of Electrical and Electronic Engineers
I/O	Input/Output
IOR	Interleaved Or Random
iRODS	Integrated Rule-Oriented Data System
ISC	International Supercomputing Conference; European equivalent to the US based SC conference. Held annually in Germany.
JSC	Jülich Supercomputing Centre (FZJ, Germany)
KB	Kilo (= 2^{10} ~10^3) Bytes (= 8 bits), also KByte
KTH	Kungliga Tekniska Högskolan (represented in PRACE by SNIC, Sweden)
LINPACK	Software library for Linear Algebra
Little-endian	Little-endian is the ordering where the least significant byte in a value is stored first (at the lowest storage address)
LiU	Linköpings universitet (represented in PRACE by SNIC, Sweden)
LLNL	Lawrence Livermore National Laboratory, Livermore, California (USA)
LRZ	Leibniz Supercomputing Centre (Garching, Germany)
LTFS	Linear Tape File System
LTO	Linear Tape-Open
MB	Mega (= 2^{20} ~ 10^6) Bytes (= 8 bits), also MByte
MB/s	Mega (= 10^6) Bytes (= 8 bits) per second, also MByte/s
MFlop/s	Mega (= 10^6) Floating point operations (usually in 64-bit, i.e. DP) per second, also MF/s
MHz	Mega (= 10^6) Hertz, frequency =10^6 periods or clock cycles per second
MIPS	Originally Microprocessor without Interlocked Pipeline Stages; a RISC processor architecture developed by MIPS Technology
MMU	Memory Management Unit
Mop/s	Mega (= 10^6) operations per second (usually integer or logic operations)
MoU	Memorandum of Understanding.
MPI	Message Passing Interface
MPP	Massively Parallel Processing (or Processor)
NCI	National Computational Infrastructure (Canberra, Australia)
NCSA	National Center for Supercomputing Applications (Illinois, USA)
NDA	Non-Disclosure Agreement. Typically signed between vendors and customers working together on products prior to their general availability or announcement.
NFS	Network File System
NIC	Network Interface Controller

NREL	National Renewable Energy Laboratory
NUMA	Non-Uniform Memory Access or Architecture
NVRAM	Non-Volatile Random Access Memory
OpenCL	Open Computing Language
OpenGL	Open Graphic Library
OpenMP	Open Multi-Processing
OS	Operating System
PCIe	Peripheral Component Interconnect express, also PCI-Express
PGAS	Partitioned Global Address Space
pNFS	Parallel Network File System
POSIX	Portable OS Interface for Unix
PRACE	Partnership for Advanced Computing in Europe; Project Acronym
PSNC	Poznan Supercomputing and Networking Centre (Poland)
QDR	Quad Data Rate
R&D&I	Research, Development and Innovation
RAM	Random Access Memory
RDMA	Remote Data Memory Access
REST	Representational State Transfer
RISC	Reduce Instruction Set Computer
RPM	Revolution per Minute
SAN	Storage Area Network
SARA	Stichting Academisch Rekencentrum Amsterdam (Netherlands)
SAS	Serial Attached SCSI
SATA	Serial Advanced Technology Attachment (bus)
SDRAM	Synchronous Dynamic Random Access Memory
SGI	Silicon Graphics, Inc.
SKU	Stock Keeping Unit
SM	Streaming Multiprocessor, also Subnet Manager
SMP	Symmetric MultiProcessing
SMR	Shingled Magnetic Recording
SNIC	Swedish National Infrastructure for Computing (Sweden)
SP	Single Precision, usually 32-bit floating point numbers
SSD	Solid State Disk or Drive
STFC	Science and Technology Facilities Council (represented in PRACE by EPSRC, United Kingdom)
SURFsara	Dutch national High Performance Computing & e-Science Support Center
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.
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
TUWIEN	Technische Universität Wien
VLIW	Very Long Instruction Word
WUE	Water Usage Effectiveness

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.

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.

D5.3 continues the work done for D5.1 [1] and D5.2 [2], updating analysis and recommendations with regard to recent developments. Material on current status of HPC site infrastructure and future plans is taken from site presentations at the 5th European HPC Centre Infrastructure workshop held in Paris in April 2014. Information gathered by work package members at SC13 (USA) and ISC'14 (Germany), the two main HPC conferences, has been used for analysis of current and near-future systems and vendor plans.

Data centre technologies for cooling and measuring efficiency metrics are covered along with optimization of applications for power efficiency. Analysis of the Top500 and Green500 lists gives an overview of recent developments of petascale systems, while the chapter on exascale and big data looks a bit more into the future and how that will affect HPC. A review of PRACE and the European HPC ecosystem in a global context provides a higher-level perspective.

1 Introduction

This is the third and final deliverable from WP5 and continues the work done for D5.2 [2] – *Best Practices for HPC Procurement and Infrastructure* – and updates it with recent developments in the last year. It contains a lot of technical detail, and is intended for persons actively working in the HPC field. Practitioners should read this document to get an overview of developments on the infrastructure side, and how it may affect planning for future data centres and systems.

The deliverable is organised into 5 main chapters. In addition to this introduction (Chapter 1) and the conclusions (Chapter 7) it contains:

- Chapter 2 – Data Centre Facilities Ecosystem – a summary of information presented at the 5th European Workshop on HPC Centre Infrastructures, hosted by CEA outside of Paris in early April 2014. Sites in Europe, USA and Australia all presented the state of their infrastructure.
- Chapter 3 – Energy efficiency in HPC – description of the trends in cooling technologies for high density IT systems, their scalability and efficiency, metrics for monitoring systems and tools for tuning application power efficiency.
- Chapter 4 – Assessment of Petascale Systems – provides an analysis of what has changed between the Top500 list in June 2013 and June 2014. The chapter is also summarising the Green500 list and the new HPCG benchmark that might provide new insights on the effectiveness of HPC systems. The business analysis section examines both data archiving and the activities of several major HPC vendors.
- Chapter 5 – Exascalability and Big Data in HPC – presents trends and basic research towards the exascale systems and data centres which will be able to support and maintain this technology.
- Chapter 6 – PRACE and the European HPC Ecosystem in a Global Context – reviews the different groups in Europe working on advancing the state of the art in HPC and how they relate to similar groupings on other continents.

Members of WP5 attended the two main HPC conferences, SC13 in Denver (USA) and ISC'14 in Leipzig (Germany), which have taken place in the time frame between D5.2 [2] and this deliverable. At those conferences information was collected on a wide number of topics related to the deliverable. Two face-to-face meetings were held to disseminate this information among the members of WP5 in preparation for this deliverable.

2 Data Centre Facilities Ecosystem

The series of European Workshops on HPC Centre Infrastructures is now well established. PRACE-PP then PRACE-1IP and PRACE-2IP gave the opportunity to accompany the consolidation of these workshops, started in 2009 at the initiative of CSCS (Switzerland), CEA (France) and LRZ (Germany).

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

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

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

The Fourth Workshop took place in April 2013, near Lugano (Switzerland), organised and hosted by CSCS with the usual CEA and LRZ support – executed during PRACE-2IP and with 60 participants.

The Fifth Workshop was also prepared and executed during PRACE-2IP, with strong involvement of WP5 members and extra PRACE sponsorship (complementing usual sponsorship brought by CEA, CSCS and LRZ – this latter one actually endorsed by GCS – Gauss Centre for Supercomputing). It was organised in April 2014 with the same structure as all the previous workshops; two days of plenary sessions followed by a half-day PRACE closed session – for PRACE sites only.

We still have the strong core of regular Workshops attendees from PRACE and non-PRACE sites but also from the technology side, i.e. providers of both IT and technical equipment. This successfully continues the creation of a sustainable joint interest group and shows a clear community response to the importance of the issues tackled by the workshops. A recent and very positive trend has been the confirmed interest of the Energy Efficient HPC Working Group (http://eehpcwg.lbl.gov/), an influential and very active initiative supported by US DOE; the EE HPC WG had representatives and active US members joining the workshop for the second time this year, expressing strong interest for continuation of participation and further collaboration.

Another evolution was the Programme Committee extension with the participation of Spain (BSC), Sweden (SNIC/KTH) and Poland (PSNC), all three active PRACE-2IP WP5 partners, in addition to the "historical" founders CEA, CSCS and LRZ – the extended Programme Committee was thus as follows:

- Guillermo Aguirre, BSC – Spain
- Ladina Gilly, CSCS – Switzerland
- Herbert Huber, LRZ – Germany
- Norbert Meyer, PSNC – Poland
- Jean-Philippe Nominé, CEA – France
- François Robin, CEA – France
- Gert Svensson, SNIC-KTH – Sweden

This year the programme mixed a number of talks (see the full program in section 8.1) with two panels fostering interaction between panellists and the audience:

- **Direct liquid cooling panel: large deployments and vendors perspectives**
  The purpose of this panel was to get feedback from vendors who have installed large configurations in production conditions.
- **European HPC R&D towards energy-efficient exascale systems**
  This more prospective panel mixed research teams and companies taking part in European funded projects.

63 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 – 38 out of these 63 participants came from PRACE member organisations. The PRACE closed session gathered 20 participants from European sites.

Fifteen (15) countries were represented, 13 from EU plus USA and Australia.

An excellent representation of PRACE sites: BSC, CEA, CINECA, FZJ, LRZ, HLRS, CSCS, STFC, SURFsara, PSNC, SNIC/KTH, SNIC/LiU, VSB-TUO, GRNET, IDRIS, CINES and NCSA.

Other European sites represented were TU Wien, U. Ulm, TU Dresden, DKRZ, AWE, Météo France, EPFL. US sites present this year and giving plenary talks were NREL and NCSA; NCI from Canberra (Australia) also gave a plenary talk.

Vendors ranged from large companies such as Bull, Cray, HP, Intel, SGI, to smaller European integrators Eurotech and Megware, as well as OVH (large service provider with their own server integration approach).

This workshop provided most of the material summarized in this chapter and in chapter 0.

2.1 Overview of HPC Facilities in Europe – Tier-0 Sites

These are among the sites that host PRACE Tier-0 resources. Currently there are six Tier-0 resources available deployed by 4 Hosting Members: France, Germany, Italy and Spain.

2.1.1 *Barcelona Supercomputing Center (BSC, Spain)*

The Barcelona Supercomputing Center is a Tier-0 site in PRACE and a hosting partner in the PRACE AISBL consortium. The MareNostrum 2 system was scaled up to 94 TFlop/s installed in 44 racks, consuming 750 kW. The machine was installed in 2006 and occupied 120 m^2 in the BSC chapel's data centre.

The electrical parameters of MareNostrum 2 are the following:

- Required: 1.25 MW, where 750 kW are for IT systems and 500 kW for the rest of the infrastructure.
- The capacity of transformers is 3 MW (2 + 1 redundancy, 3 x 1 MW transformers).
- Increased electrical reliability was done by 180 kW UPS and a 360 kW generator (in total 540 kW).

The cooling system requires additional 1 MW of energy.

The MareNostrum 3 system installed in 2012 has a peak performance of 1.1 PFlop/s and consists of 52 racks placed in the space previously used by MareNostrum 2.

The requirements of electrical energy after 6 years are following:

- Installed 5 MW of input power (2+1 redundancy, 2x2 MW + 1x1 MW transformers).
- The facility consumes 1.6 MW, where 1.1 MW is used for IT and 500 kW for everything else.
- Redundant power of 540 kW is provided, 180 kW from a UPS and 360 kW from a generator.

The cooling system requires 1.3 MW of energy.

MareNostrum 3 is cooled by rear doors heat exchanger. The BMS (Building Management System) is performing an automatic rotation of redundant components and temperature control of water circuits and room air.

The new BSC headquarter (under construction) will include a new data centre (815 m^2 x 6,5 m high) and offices.

2.1.2 *Commissariat à l'Énergie Atomique aux Énergies Alternatives (CEA, France)*

The French Tier0 system called CURIE, funded by GENCI as Hosting Member is located and operated by CEA on the Très Grand Centre de calcul du CEA (TGCC), at Bruyères-le-Chatel just south of Paris. Computer facilities hosted and owned by CEA are split into two parts, one civilian and one military. Curie is hosted at TGCC along with several other civilian systems

including the supercomputer Airain shared between CEA and industrial partners. Located "inside the fences", a few hundred meters from the TGCC, the classified TERA facility hosts classified research and the TERA-100 system, a sister system of Curie. A summary of the systems hosted or owned by CEA can be found in Table 1.

Teratec, a HPC technology park is also situated closely to TGCC and provides office space and lab areas for HPC companies.

TGCC has 2,600 m^2 floor space for computer rooms, and the capacity to handle 60 MW of electrical power. CEA has been collaborating with energy company EDF since 2012 to use the MAPE software for monitoring power consumption, cooling and ventilation. Optimizations based on these measurements have led to a reduction in operational costs by 4% and a PUE just above 1.3 for TERA. After the successful use in TERA it was also installed in the TGCC, where it has lowered the PUE from 1.55 to 1.5 and reduced the yearly cost by 60 k€. Due to the current under-utilization of TGCC, the PUE is expected to be lower after Curie-2 is installed. Heat generated by the upcoming TERA-1000 (expected to produce around 5 MW) will be used to heat the offices of 2000 employees in the surrounding office buildings.

Bull is the vendor that has delivered the three major compute clusters hosted by CEA. Bullx B510 thin nodes and S6010 fat nodes are used in both Curie and Airain. These two systems also share 10 PB of disk space in addition to each systems private disk space.

System	Peak Performance	Disk space	Compute nodes (cores, RAM) (Bullx)
Curie	2 PFlop/s	5 PB	90x quad-S6010 (128 NH, 512 GB) 5040 x B510 (16 SB, 64 GB) 144 x B505 (2 x M2090 each)
TERA-100	1.25 PFlop/s	5 PB	S6010/S6030
Airain	0.4 PFlop/s	2.3 PB	360 x B510 (20 IVB, 64 GB) 594 x B510 (16 SB, 64 GB) 180 x B510 (16 SB, 128 GB) [FG]

Table 1: Current systems hosted or owned by CEA

Curie is used for public research, with 80% of the capacity allocated as Tier-0 for PRACE and the remaining 20% going to national research projects. TERA-100 is wholly dedicated to classified CEA/DAM computations.

Airain is being used for more industrial purposes, and is the compute resource for Centre de Calcul Recherche et Technologie (CCRT). This is a consortium of companies that have pooled their resources into a system for large computations. CCRT also provides dedicated resources for the bioinformatics project France Génomique (FG). These resources include compute nodes, short-term storage on disk and long term tape archiving.

Lustre storage is provided on DDN hardware, and CEA is one of the large contributors to the development of HSM functionality in Lustre. This is for example used to automatically move data generated by France Génomique on the Airain system to long-term archive storage. Tape libraries from Bull and StorageTek are used for backups and archiving.

Scheduling system Slurm is another open source software that CEA is doing development on.

2.1.3 *Höchstleistungsrechenzentrum Stuttgart (HLRS, Germany)*

HLRS is a federal HPC centre and a Stuttgart University Institute. The users are both scientific users and industrial users (through a specific structure hww GmbH[1]).

The infrastructure includes two building: the "old" hall (200 m^2, max. 200 kW water and air cooling), and the "new" hall upgraded in 2011 (800 m^2, max 1000 kW air cooling and 4000 kW water cooling).

The current system is a CRAY XE6 ("Hermit") with a peak performance of 1 PFlop/s and a maximum power consumption of 2 MW (1.6 MW typical).

A new system, CRAY XC30 ("Hornet") will be installed in Q3-2014 and will run in parallel with Hermit. This new system, based on Intel Haswell processors, will have a peak performance of 3.8 TFlop/s and a maximum power consumption of 1.8 MW (1.4 MW typical).

In Q3-2015, "Hermit" will be replaced by an extension of "Hornet" with 20 new cabinets (in addition to the 21 racks installed in Q3-2014). The total system will then have a peak performance of 7.4 PFlop/s and a maximum power consumption of 3.5 MW (2.8 MW typical).

A new project "Sustainability in HPC centers" just started at HLRS. It aims at defining and implementing a sustainability strategy based on the analysis of all aspects having an environmental impact. The EMAS certification (Eco Management and Audit Scheme) is targeted.

2.1.4 *Leibniz-Rechenzentrum (LRZ, Germany)*

LRZ is a member of the German Gauss Supercomputing Centre and is also hosting one of the PRACE Tier-0 systems, SuperMUC. It is also involved in research into future HPC systems, with one of the focus areas being reducing emissions caused by HPC. Their four pillar model[2] for optimizing the data centre includes building infrastructure, system hardware, system software and applications.

Current data centre facility has 6,393.5 m^2 of floor space for infrastructure, of which 3,160.5 m^2 is for IT equipment. The floor space is split into six rooms on three different floors. Power is provided via 2 x 10 MW 20 kV connections.

SuperMUC is the largest system installed at LRZ, and has a peak performance of 3.2 PFlop/s. Expansion with phase 2 is planned, and will add another 3.2 PFlop/s to the system. 90% of the heat generated by SuperMUC is captured by water cooling.

Hydraulic gates were installed when building the water cooling infrastructure for SuperMUC, since that had been used in an older installation. However, this turns out to be unnecessary with regards to the flow of water, while it leads to approx. 5% loss in the cooling loop. Therefore, such a device should be avoided.

The heat being generated by the systems is being reused to heat the office buildings. However, the current amount of waste heat could heat 26,000 m^2 of office space. That is ten times the capacity of nearby offices (like other countries with snowy winters, German buildings are well insulated). Another option for the re-use of waste heat being explored with the CoolMUC prototype is feeding it into an adsorption chiller to generate cooling.

[1] Höchstleistungsrechner für Wissenschaft und Wirtschaft – http://www.hww.de/en/
[2] https://www.lrz.de/wir/green-it_en/

Energy to solution is a concept used by LRZ. Jobs are run in a test mode to find out the optimal processor frequency for the application and type of input data. Most applications have their sweet spot at 2.3 GHz on SuperMUC. Users may tag their jobs with a name, which indicates which type of job it is. The same binary combined with the same type of input should always be using the same tag. LRZ has being working together with IBM to include this functionality into LoadLeveler (LL). Load Sharing Facility (LSF) will in the future also include this functionality, since it is the intended LL replacement from IBM.

Energy aware scheduling is estimated to save 160 k€/year in reduced energy consumption. Currently the energy consumption is costing more than 400 k€/month. Future plans are to port some energy aware scheduling features to the Slurm batch scheduler.

2.2 Overview of HPC Facilities Projects in Europe – Tier-1 Sites

PRACE Tier-1 sites are hosting national resources where parts of the core hours are dedicated to projects getting access via the Tier-1 calls (also known as DECI calls). A list of all Tier-1 resources is available[3] at the PRACE website.

2.2.1 *Swiss National Supercomputing Centre (CSCS, Switzerland)*

CSCS (Centro Svizzero di Calcolo Scientifico) is located in Lugano, in a new facility that started operation in March 2012. “Piz Daint”, the fastest European system is hosted in this facility. The site is located 3 km from Lake Lugano, from where the cooling water for the data centre is drawn.

Energy efficiency was a key design criterion when the data centre was built, targeting a PUE below 1.25. For this reason, UPS protection is only provided for critical infrastructure, although there is space available for adding rotary UPS if the incoming power quality is degraded in the future. The location of the building, right next to the Lugano fire department also enabled CSCS to get a unique exception from an extinguishing system.

CSCS is very satisfied with the new data centre and its very flexible IT floor space with no pillars hindering the placement of systems. Having a full height facility level between the data centre floor and the basement has enabled major installation work to be done without affecting the IT operations.

Energy measurements are taken at many points, which was part of the design of the building. These measurements, and a separate software tool for monitoring and managing the electrical network, were instrumental in achieving Level 3 measurements for Green500 submissions. The focus on energy efficiency has created an environment that brings together staff from facility management, IT and research, and has proven to be an excellent way of increasing the understanding and appreciation for each others work.

Current IT load is around 3 MW with a PUE of 1.23. A small turbine is being planned to be fitted in the return water pipe, to produce power from the flow of currently 80-100 l/s.

The cooling system based on the extraction of water from -45m in Lake Lugano has proved to be very energy efficient, but has also brought up challenges related to the biological and chemical reactions at this depth. Currently the bacteria Leptothrix Ochracea is clogging the inlets, and also fills the filters and heat exchangers in the data centre, requiring constant attention. A closed water loop may have avoided this particular problem, but would have been fraught with its own set of issues. Despite this issue, CSCS would repeat this design decision.

[3] http://www.prace-ri.eu/Tier-1-Resources?lang=en

Stray currents in pump ball bearings have also been an issue, damaging the bearings of some of the pumps. This is a known issue with variable frequency drives. Switching to ceramic ball bearings reduced the problem and further work has been done to better insulate pump parts to avoid the stray currents forming in the first place.

2.2.2 *Poznan Supercomputing and Networking Center (PSNC, Poland)*

PSNC (Poznan Supercomputing and Networking Center) has been experimenting a prototype of liquid cooled system provided by Iceotope (company based in UK), in the context of PRACE-1/2IP project.

This prototype is composed of 40 nodes, including non-GPU and GPU nodes with a QDR IB interconnect. Each node is an immersive cooling enclosure that uses 3M Novec fluid for cooling the components. Each blade includes a hot plate (heat exchanger) in order to cool the internal liquid with water which remains inside the rack. This water is cooled by the separate water-glycol loop by rack level heat exchangers.

At PNSC the cooling is provided by heat exchangers on the roof with adiabatic support in case of high outside temperature. An emergency connection to the facility chilled water loop is provided to handle downtime of the primary coolant loop.

A strong focus was put at PSNC for collecting data for both cooling and server parameters. The PUE is around 1.02.

Several potential issues with DLC system were mentioned during the presentation:

- Monitoring of the coolant water is critical. A leak in a PDU caused a degradation of the water quality in the loop inside the rack. The consequence was that the whole loop was destroyed by electrolysis in 3 weeks and it took 2 months to replace all the components.
- The pumps use most of the energy. Therefore pumps must run at the lowest possible speed.
- One side effect of having a W3/W4 cooling in a computer room is that the temperature of the servers themselves are increased; therefore the air cooling system of the room has to handle the waste heat. In the worst case the DLC system may radiate up to 15% of the heat to the ambient air.
- Some oscillations may occur in the cooling system especially when inlet temperature for cooling gets higher. Using buffer tanks may reduce this issue.
- Inlet temperature in the range of 25°C-30°C is good (25°C being preferable) in terms of energy efficiency. It makes cooling without chillers possible most of the time (over 98% of hours per year).

Future work includes reduction of processor frequency during periods of high outside temperature in order to further avoid the usage of chillers.

2.2.3 *Science and Technology Facilities Council (STFC, UK)*

STFC (Science and Technology Facilities Council) is one of the seven research councils in UK. The role of STFC is to plan, develop and operate large scale facilities needed for research in the UK. The Scientific Computing Department (170 staff) supports over 7500 users and operates several large supercomputers including iDataPlex and BlueGene systems.

The Hartree Centre, established close to Daresbury Laboratory, aims at helping UK public companies exploit scientific computing. It focuses on industry and makes it possible for public companies to access skills and resources (compute, storage) mainly located at Daresbury Laboratory and at Rutherford Appleton Laboratory.

The EECR (Energy Efficient Computing Research) initiative at Hartree Centre was launched in early 2013 in order "to firmly establish the UK as the world leader in energy efficient supercomputer software development to meet big data challenges." One motivation is the growth of electricity consumption for ICT, which in the UK is expected to account for 20% of the total electricity consumption by 2020.

EECR is based on three strands: datacenters, computers and peripherals, software and algorithms. Improvement of existing technologies, as well as opportunities of new technologies (low power processors, data flow/FPGA and novel cooling) will be investigated.

Partners and solution that will be procured include:

- IBM
  - NeXtScale, iDataPlex, Intel Phi (£5M)
  - BG Active Storage (£5M)
  - Low Power Processors (£0.7M) - NeXtScale + ARM CPUs
  - Big Data & Data Analytics (£4M)
- Insight/Clustervision
  - Novel Cooling Technology (£1M)
- Viglen/Maxeler
  - Dataflow Technology (£1M)

In addition, work will be conducted with Concurrent Thinking limited on techniques for energy efficient datacenters.

The operation is planned to start in July 2014 and the first results are expected in 2015.

2.2.4 *SURFsara (Netherlands)*

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

The current large system operated by SURFsara (Bull) was installed in 2013. It is installed in the SURFsara data centre in Amsterdam. This system, currently at 270 TFlop/s, will be upgraded to approximately 1.5 PFlop/s by the end of 2014.

Although this supercomputer upgrade will be performed in the current data centre, it touches on its limits in terms of space, power and efficiency (PUE). Since this data centre can't be expanded by SURFsara both for technical reasons and legal reasons (the data centre is now run by Vancis, a different company), two options were possible for SURFsara: either build a new HPC data centre of its own, or start a tender to rent space in a data centre that meets SURFsara's requirements for the near and longer term future.

In order to investigate both options, SURFsara defined the requirements for a new data centre including location (near SURFsara location in Amsterdam Science Park), flexibility in terms of space, power and cooling, good PUE, warm water cooling (W3), floor strength, projected needs in terms of number, type, power consumption of racks. On the one hand flexibility is crucial to being able to innovate and replace an existing supercomputing facility with the new state of the art technology. On the other hand flexibility is expensive. In the tender SURFsara dealt with this tension by specifying growth scenarios with bounded annual flexibility.

Based on these requirements, it turned out that, with the volume and growth path envisaged by SURFsara, renting space in commercial data centre was possible and would be 25% to 30% cheaper than building a data centre. SURFsara signed a contract with TeleCity for two

floors of a 15-floor commercial data centre, plus a specially conditioned room for a tape library on another floor. The offer from TeleCity entailed that SURFsara can have its own "data centre within a data centre". The offer was compatible with the requirements of SURFsara with a good provision for extension. In addition, TeleCity is willing to be a partner in exploring new technologies with SURFsara at the system level.

2.2.5 *VŠB–Technical University of Ostrava (VSB-TUO, Czech Republic)*

VŠB-TUO hosts the only centre of excellence in the IT field in the Czech Republic, IT4Innovation.

The current supercomputer is a cluster with a performance of 75 TFlop/s (LINPACK). It is installed in a mobile data centre and based on a container provided by Bull (MOBULL). The supercomputer is in production since August 2013. It includes:

- 180 thin nodes (dual Sandy Bridge, 64 GB)
- 23 GPU accelerated nodes (NVidia K20)
- MIC accelerated nodes (Intel Xeon Phi)
- 2 fat nodes with large memory (512 GB).

The total power consumption is 74 kW with a PUE of 1.38.

The experience with the supercomputer is one of good stability and reliability, except for the hard drives in the compute nodes (7.1% failed). The local support from Bull is very responsive. From a system software point of view, several issues were listed including poor design of HA features, slow adoption of upstream releases (leading to security risks) and painful upgrade.

In the future there will be a new computer centre (summer 2014) and a new cluster (procurement on-going, target performance 1 PFlop/s LINPACK).

2.3 Overview of HPC Facilities Projects in Europe – Other sites

Sites in this section are not hosting a PRACE resource, although they may be involved in PRACE in some other capacity.

2.3.1 *Atomic Weapons Establishment (AWE, UK)*

AWE (Atomic Weapon Establishment) is in charge of design, manufacturing and certification of UK nuclear weapons. Over the last 5 years, AWE has made many changes to its IT infrastructure with the consolidation of corporate data centres (from 7 to 2) and HPC DCs (from 3 to 2).

Regarding corporate DCs, the systems are cooled by water, with an inlet temperature of 25°C. Because of density issues, a future refresh of the cooling system will be necessary.

Two HPC DCs are in operation:

- "Small DC", used for resilience of data and continuity of a small HPC service,
- "Old DC" divided into 3 halls (400 kW, 100 kW and 1 MW); this "old DC" will be decommissioned in the future.

A new HPC DC is almost finished. The project started in 2010 and the installation of the new Petascale supercomputer ordered by AWE from SGI started in January 2014. Final handover of the building is due June 2014. This DC is located in a legacy facility and will include office space. It is designed in order to provide flexible power and cooling capacity, currently with 15-19°C water and possibly with 24°C water later. The power supply is protected by UPS and

diesel generators (n+1). A PUE below 1.2 is expected at the beginning, it should be below 1.1 with warmer water. The computer room floor space is 500 m^2 for 2.5 MW IT load at the beginning; the room floor space will be increased to 1,000 m^2 and the power to 9.5 MW IT load split across the 2 halls in the future.

2.3.2 *Centre Informatique National de l'Enseignement Supérieur (CINES, France)*

CINES is the French national computing centre for higher education and universities. It provides the French public research community with computing resources and services. It is located in Montpellier with a staff of around 60 people. CINES also has a mission of national servers hosting and a mission of data archiving and long term preservation for universities and public research organizations.

A new computer room (SM5) will be available in June 2014 for hosting a new computer expected in this time frame. It will complement the existing 4 computer rooms (total floor space of 820m^2) that are difficult to use because of a circular shape. This room is divided into three areas:

- Storage area (130 m^2) designed for air cooling.
- Computing area (370 m^2) designed for "hot" water cooling and air cooling.
- Logistic area (120 m^2) for unpacking and preparation of parts.

The cooling capacity is 2 MW for "hot" water, 1.4 MW for "cold" water.

Two other evolutions of the site were presented:

- Update of the electrical network with two primary power sources (2.6 MW and 10 MW), UPS (10 minutes full backup) and two local generators. The improvement of the redundancy of the distribution is on the way.
- New "hot" water loop that should make it possible to achieve under the worst meteorological case, 30°C at the CDU (Coolant Distribution Unit) level with adiabatic coolers (with humidification). A tight monitoring and watch is planned for the coolers because of the risk related to the obstruction of radiation by bugs or pollen.

2.3.3 *Greek Research and Technology Network (GRNET, Greece)*

GRNET (The Greek Research and Technology Network) is in the process of deploying a supercomputer. The call for tender was issued on September 2013 with a budget of 2.6 M€ including IT equipment and infrastructure. A contract has been signed with IBM in July 2014. The system is expected to be operational by the end of 2014.

This signed contract includes:

- An HPC system based on IBM's NeXtScale platform, including a total of 426 compute Intel ® Xeon ® E5-2680 v2 processors (Ivy Bridge) nodes, interconnected via a non blocking IB FDR network, that will offer more than 8,500 processor cores (CPU cores) and a performance of around 178 TFlop/s (LINPACK).
- A parallel file system, based on GPFS, with a capacity close to 1 PB (raw) and a bandwidth of 10 GB/s.
- Service and support for 5 years for the IT equipment.
- Infrastructure equipment, suited for the power consumption of the IT equipment (185 kW at full utilization, 175 kW during LINPACK), for cooling (in-row cooling units with hot aisle containment and chillers) and for power distribution in the computer room.
- Infrastructure equipment support and maintenance services for 7 years.

The procurement process was made with a strong focus on benchmarks (60% of the overall technical evaluation which accounts for 80% of the overall evaluation).

A second tender, currently at its final stages, was made for infrastructure equipment not included in the supercomputer tender. This includes:

- Construction works
- Power supply
  - New 1600 kVA MV/LN transformer
  - Central Electrical panels
  - UPS + batteries (400kW) (all IT equipment are powered through UPS because of the low quality of electricity supply in the area were the computer centre is installed)
- Fire detection and suppression
- Access control, alarm, CCTV systems
- Power and environmental monitoring

An extension of IT equipment (possibly fat nodes, accelerators, storage or thin nodes) and associated infrastructure is planned for next year. A RFP, based on user requirements under analysis, is planned for the end of this year.

2.3.4 *Météo France (France)*

Météo France is the French national weather service. In 2012, they adopted a four year strategic plan with the French government. This includes a HPC roadmap leading to a total compute power of 2 x 2.85 PFlop/s by the end of 2016.

The two systems are hosted in separate data centres. Previous infrastructure, especially data-handling systems had to be maintained, and they will remain in the CNC (Centre National de Calcul). A new computing room was delivered in November 2013, part of the new ECA (Espace Clément Ader) building. Having two locations will allow fulfilling a high level of reliability and evolution flexibility.

ECA offers a surface of 524 m² and 1 MW of IT power, water temperature range is W1 (13-17°C), it has been decided and designed 2 years ago. PUE and ERE are still under measurements. PUE for CNC is 1.7. A new project has been started for optimizing the PUE of both machine rooms.

The roadmap for Météo France resources can be seen in Table 2:

CNC machine room	**522 TFlop/s** Peak 56 racks Bullx DLC 1,008 compute nodes Intel Ivy Bridge EP Fat Tree IB FDR Lustre 2PB, 69 GB/s Storage bullx SCS 209 TB **Available 09/2013**	**2.85 PFlop/s** Peak 56+45 Racks Bullx DLC 1,800 compute nodes Intel Broadwell EP Fat Tree IB FDR Lustre 3.57PB, 138 GB/s Storage Bullx SCS 400TB **Available 06/2016**

ECA machine room	**513 TFlop/s** Peak 55 racks Bullx DLC 990 compute nodes Intel Ivy Bridge EP Fat Tree IB FDR Lustre 1.53PB, 46GB/s Storage Bullx SCS 135 TB **Available 03/2014**	**2.85 PFlop/s** Peak 55+45 Racks Bullx DLC 1,800 compute nodes Intel Broadwell EP Fat Tree IB FDR Lustre 2.55PB, 92 GB/s Storage Bullx SCS 135TB **Available 06/2016**

Table 2: Current and planned Météo France systems

2.3.5 *Nationellt superdatorcentrum (NSC, Sweden)*

NSC-LiU (National Supercomputer Centre-Linköping University) provides HPC services to academic institutions in the country as well as to partners SMHI (weather forecast) and SAAB, since 1989, with a staff of 30 people. The talk focused on major evolutions since the presentation given during the previous workshop.

In terms of building, the "Bunkern" building is no longer used. Half of the systems hosted in this building were decommissioned; the other systems were moved to the "Hangaren" building (250 m^2).

The utilization of parts of the new "Kärnhuset" building (one cell, 300 m^2, no raised floor) started after the acceptance test of the cooling system. This building is flexible, with room for expansion in three empty cells. Two weak points regarding this building were listed: entrance system not working as designed, and the UPS delivery (reseller not knowledgeable).

The first system moved from the "Hangaren" building to this building, is Triolith, the largest system with a peak performance of close to 450 TFlop/s. The cell hosts also new disk storage (IBM GSS) with 4 PB of raw space. The total power consumption is 480 kW with a PUE of 1.07 including 10 kW for fans and pumps and 20 kW for UPS loses (excluding the generation of the chilled water).

The power supply is provided through UPS (six UPS of 200 kVA including one for fans and pumps).

The racks are cooled by air with hot air containment. The cooling is provided in winter by cooling towers and in summer by absorption chillers powered by heat from the district heating network. A buffer tank of 30 m^3 can provide cooling for 30 minutes with 1 MW IT load.

The next system is a joint Swedish and Norwegian Met system. An open procurement is going on at this time for a total budget of 32 MSEK (≈3.5 M€) including operation for 4 years. It will be located in the cell of the "Kärnhuset" building already used.

2.3.6 *OVH (OVH, France)*

OVH is a European infrastructure and service provider with 15 data centres located worldwide, 700,000 customers and 18 million hosted applications. The company has successfully completed a global datacentre concept, maintaining full control of variables and demonstrating strong expertise in high availability, security and energy efficiency.

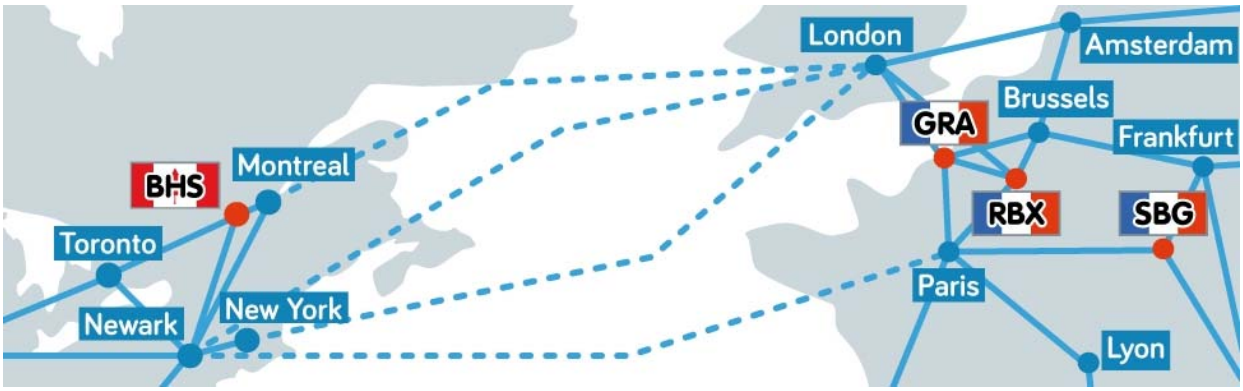
The physical security is achieved by:

- OVH data centres are strictly for OVH usage, servers can only be physically accessed by authorised employees.

- Access restricted by security badge control system, video surveillance and security personnel 24/7 on-site.
- Rooms fitted with smoke detection systems.
- Technicians on site 24/7.

The OVH datacentres are powered by two separate electrical power supplies and are also equipped with UPS devices and power generators able to provide the full power needed for the data centre. Power generators have an initial autonomy of 48 hours to counteract any failure of the electricity supply network

As already mentioned, OVH has a total of 15 datacentres: 13 located in France in Roubaix, Gravelines, Paris and Strasbourg on sites which are over 200 km apart, and 2 located in Canada.



This enables OVH to guarantee continuity of service, even in the event of a major incident. For example, the backup of web hosting data located in Paris is replicated every day on the Roubaix machines, while hubiC data (online storage platform) is duplicated on three sites. OVH also offers dedicated server customers the option of carrying out data backups across multiple datacentres.

OVH data centres are equipped with the newest cooling technology:

- 98% of OVH hosting rooms are free from air conditioning.
- Water cooling enables 70% of heat emitted by the processors to be dispersed.
- Free cooling evacuates the remaining 30%.
- Energy costs halved.
- The best value of PUE parameter is 1.09.

OVH is certified ISO 27001:2005 for providing and operating dedicated cloud computing infrastructures. OVH is based on the ISO 27002 and ISO 27005 security management and risk assessment norms and associated processes. OVH has obtained SOC 1 and 2 type I certification for 3 datacentres in France and 1 in Canada, which certifies the security level for OVH Dedicated Cloud.

2.3.7 *Technische Universität Wien (TU Wien, Austria)*

VSC is a consortium composed of the main Austrian universities, servicing the HPC needs of a large and heterogeneous community. During the last three technological renewal cycles (2009-2014) VSC facilities operated the transition from air cooling (hot aisle and Knurr CoolLoop) to immersion cooling, with traditional cluster architecture (without accelerators).

For the last cycle (the third) a new machine room was built adapting an old research lab facility (built around 1970). During 2011 VSC decided to test a prototype installation with

immersion cooling (oil based) from which they gained a good feedback. They tested several different systems "in oil" with the only main issue in power supply defects, promptly fixed.

The gained experience in the field was used to determine a model to estimate TCO, and then to plan the call of tender for the aforementioned third cycle system. Given the plan for a new room, it was decided to adapt the infrastructure for traditional air or direct/indirect liquid cooling, and immersion cooling, with raised floor and oil-proof containment coating underneath. The tender didn't specify the cooling technology, but vendors were asked to supply everything except cooling water, including heat exchangers. VSC also expressed the preference for higher inlet temperature technology.

The offer that best fitted the TCO-based criteria was for an "oil cooled" system, with inlet temperature of 43°C, with water/glycol-mixture exchangers and free coolers. The full system configuration consists of 2,208 compute nodes distributed among 23 racks (custom made) and is capable of around 600 TFlop/s. Despite the immersion technology, the total power dissipated by each rack is about 23 kW.

The VSC technical team faced a few "not typical" challenges during tender specifications: environmental protection against oil spills; fire protection; the lack of standards and legal requirements in the field of immersion oil-cooling (the difficulty is not that the oil is particularly hazardous, but nobody seems to know precisely what the requirements are).

2.4 Overview of HPC Facilities Projects in US

This section covers US sites that held presentations about their infrastructure at the workshop.

2.4.1 *National Center for Supercomputing Applications (NCSA)*

Blue Waters is the 288 cabinets Cray XE6/XK6 hosted at the National Petascale Computing Facility at the University of Illinois, supported by the National Science Foundation and the University of Illinois. It is claimed to be the fastest supercomputer on a university campus. The total available RAM is around 1.5 Pbytes. The online storage is more than 25 Pbytes while near line hierarchical storage is more than 300 Pbytes (based on tape). The main interconnect can deliver 1.2Pbytes/s to the online storage.

The time to solution/insight is the driver of the whole project, the "preferred" metric for the Sustained Performance evaluation. The consensus of many papers/experts is that the only real, meaningful metric that can compare systems or implementations; is the time it takes to solve a defined, real problem. Work is a task to carry out or a problem to solve. Just like in the real world, work is not a rate, it is not a speed, it is a quantity. Hence a good evaluation compares how much time it takes to do an amount of meaningful (productive) work referred to as the system's potential to do the work. The cost effectiveness is then the system's potential over system's cost ratio.

Different classes of benchmarks were considered to measure the Blue Waters performance. Original NSF benchmarks at two scales, full and modest size and other applications based on the Sustained Petascale Performance (SPP) metric, that aims to expand the original requirements as it is a time to solution metric that uses the planned applications on representative parts of the science team problems: represents end to end problem run including I/O, pre and post phases, etc.; coverage for science areas, algorithmic methods, scale. The input, problem sizes, included physics, and I/O performed by each benchmark is comparable to the simulations proposed by the corresponding science team for scientific discovery. Well defined reference operation counts were used to represent work across disciplines. Each benchmark is sized to use one-fifth to one-half of the number of nodes in the

full system. At least three SPP applications run at full system size. SPP tests demonstrated the real problems can scale to more than two thousands nodes and deliver more than one PFlop/s per real problem/application. Blue Waters allowed for unprecedented results in various fields as: Computational Microscope (order of magnitude increase in number of atoms per simulation); Turbulent Stellar Hydrodynamics (trillion cell, multifluid CFD simulation); Magnetosphere; Tornadoes and parent supercells; Hurricane Sandy (3Km resolution); and many others.

The National Petascale Computing facility is a modern data centre that covers around 8,360 m^2 with a computer room of about 1,860 m^2 (2,790 m^2 total raised floor). The plant is capable of sustaining 24 MW, expandable to 100 MW, and another 4,645 m^2 of machine room, LEED certified at Gold level (PUE 1.1-1.2). The staff is participating in the Energy Efficient HPC working group. The cooling infrastructure is composed of 5 control loops: Air loop in cabinets; Freon Coolant loops (top rack); Machine room Cool water loop; facility external Cooling Towers and University Mechanical Chillers.

2.4.2 *National Renewable Energy Laboratory (NREL)*

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

The new HPC data centre in NREL's Energy Systems Integration Facility (ESIF) is built as a showcase facility with a PUE goal of 1.06. It is rated at Platinum level, based on Leadership in Energy and Environmental Design (LEED) ranking system. Holistic thinking in the whole process is claimed in order to build one of the most energy efficient data centres in the world: 20 year planning horizon; 10 MW total power capability; evaporative cooling (no mechanical); warm water liquid cooling and waste heat re-use to heat labs and offices; leveraged expertise in energy efficient buildings; integrated "chips to bricks approach".

Cooling efficiency means leveraging a better heat exchange and reduce energy for coolant transport. Liquids are 1000x more efficient than air in heat exchange and require 10x less energy for transportation (fans vs. pumps). Furthermore Liquid-to-liquid heat exchangers have a closer approach temperature than Liquid-to-air, yielding increased economizer hours (leverage favourable climate). Benefits of liquid cooling (especially hot liquid) include: better thermal stability for components, allowing for more time in "Turbo mode" and better MTBF; reduction of condensation and suppression of expensive and inefficient chillers. Nevertheless, liquid cooling needs some new considerations: PH; bacteria; dissolved solids; chemical properties; type of pipes etc. It is highly recommended to follow the latest ASHRAE specifications in order to avoid extra costs during maintenance.

The other side of efficiency is power distribution: working the whole data centre with 480 VAC eliminates the needs for intermediate conversions.

Data centre efficiency optimisation copes with a three dimensional space: IT power consumption, facility PUE and energy reuse. Reducing PUE is a well known process but in order to improve sustainability it is necessary to start considering further metrics such as ERE, CUE and WUE that better help to take into account the full three dimensional space mentioned earlier.

ESIF has now more than 14 months of operations and the "Peregrine" platform (more than thirteen thousands Xeon processors plus half thousand Xeon Phi for a total sustained of 1.19 PFlop/s) is fully operational. The challenge is in energy consumption optimization, the "Extreme Scale Energy Management": how to cope with new high priority jobs to be run without exceeding the total power consumption constraints (far less than the theoretical peak)? How to "accelerate" running jobs to place new large capability jobs? How to manage

in order to reduce peaks of energy and maintain a balanced figure toward the distribution service operator (utility pricing is based on peak demand)? The answers to these challenges is the matter for further research.

2.5 Overview of HPC Facilities Projects in Asia and Oceania

This section covers Asian and Oceanian sites that held presentations about their infrastructure at the workshop.

2.5.1 *National Computational Infrastructure (NCI, Australia)*

NCI is located in Canberra and was originally part of the only federally funded university in Australia. This also means that it is located on the university campus, and has a number of constraints on available space.

The centre employs 38 people for operations and support; a group of an additional 20 people is focusing on more research related activities. National user community being supported consists of around 3,000 users and 600 projects.

State level financing only covers investments, all operational expenses are borne by the users. This means that most of the financing comes from other organizations who can control their share of the resources. Time is allocated on a quarterly basis, and must be used before the end of the quarter.

The Australian Bureau of Meteorology is one example of a partner organisation, collaborating with NCI on weather forecasting and climate research. NCI is also operating a private cloud and data storage for research groups.

The main compute resource is Raijin, consisting of 3,592 Fujitsu compute nodes with Intel Sandy Bridge cores and 157 TB RAM, all connected with FDR-14 InfiniBand. The system consumes 1.2 MW of electricity at full load, and has a PUE of 1.2. Hot aisle containment is used in the data centre, with standardised rack sizes. However, blank panels are still needed to improve the airflow for empty spaces. NCI has been using 5 mm thick corrugated plastic instead of traditional blanking plates to reduce the installation work.

Storage available on the compute systems includes 7.6 PB disk for scratch and 7.2 + 4.1 PB disk for more long-term storage. Backups and archiving uses a Spectra Logic tape library with 12.3 PB tape storage, and a 1 TB disk drive system as a frontend.

NCI has a 915 m^2 (IT floor space) data centre with 3 MW electrical capacity, backup power provided by a 0.5 MW UPS and a 1.1 MW diesel generator. Cooling is provided by two 1.8 MW water loops. Free cooling is used when possible, using a threshold of 16.1°C this can be done 84% of the year. Measurements and testing is being done to determine the best set point for turning off the free cooling and running on chillers instead.

Being located in the middle of a university campus means that there are restrictions on how much noise can be generated, and how beautiful the building must be. This affects for example the placement of cooling equipment on the roof. Another constraint was that no significant trees could be taken down when the data centre was built. This has led to a problem with bugs, real physical bugs, and leaves that clog the strainers of the pumps for the cooling towers on the roof. This severely limited the flow rate, and increased the power consumption for the pumps by 50%.

Future plans include a 8-10 PFlop/s system in the 2016-2017 timeframe with a power budget of 4 MW.

2.6 Chapter summary and trends

The summary and trends identified one year ago (see D5.2, section 2.5) have been confirmed by the presentations made during the 5th European workshop on HPC centre infrastructures. However, it is worth noting stronger focus on some topics.

Cooling:

- Even if "warm" water cooling is the main trend, air cooling on the one hand, immersive cooling on the other hand are options worth considering.
- For "warm" water cooling, the terminology introduced in 2011 by ASHRAE (W1 to W5) should be used in order to avoid confusion between very different water inlet temperatures (W1=17°C, W5>45°C). Using high W water temperatures should be done with care since comfort in the computer room can be affected and since transfer of heat between cooling systems at different temperature in a computer room can affect the overall efficiency. In addition, water quality is very important (see ASHRAE standards) and water quality problems (pH, bacteria, dissolved solids) may have severe consequences.
- For air cooling, segregation of air flow is necessary. It can be achieved by different means, depending on the computer room. Best results need computer rooms specially designed for this purpose but simple solutions can be used in conventional computer rooms.
- For immersion cooling, interesting results are reported both in terms of energy efficiency and in terms of ease of use (including maintenance). However the adoption is slow, partially due to the lack of experience, especially in the fields of long term material tolerance and fire security.
- The interest for heat reuse for heating, but also for producing cold with adsorption chillers, is growing.

Power supply:

- The trend towards more and more variability in the power consumption of IT equipment is confirmed. This leads to several important issues including variability of the power drawn from the electrical grid and the need for the facility (cooling / power supply) to handle properly such variation. Negotiations with electricity suppliers regarding this point may in some cases lead to additional costs for the site.
- The target power consumption for exascale computers is still 20 MW, however this is not a hard limit.
- TUE introduced one year ago by the EEHPC working group is getting more and more interest from the large sites. The importance of this group is growing and, among a lot of actions, the EEHPC working group has provided a great contribution to the improvement of measurements for the Green500.

Location:

- Even if requirements for hosting large supercomputers are different from the usual requirements of system hosted in commercial data centre, it turns out that in some cases hosting a supercomputer in a commercial data centre is possible and may lead to lower costs.

3 Energy efficiency in HPC

Energy efficiency has become a very important topic in the last few years. Growing energy bills, which over the lifespan of the machine are overshadowing the acquisition costs of current supercomputers has forced a change of focus. When selecting a new machine one has to answer not only the question: "how fast is it?" but also "what will be the cost of operating the system?". Answering the question about the speed of the machine was not a simple one as it is dependent on plethora of variables. Adding another dimension, energy efficiency, has made the decision even harder as efficiency may mean different things. The Top500 list started collecting how much power is used for LINPACK runs and a derivative list: the Green500 was created to focus on this aspect. However, this only includes the power used by compute nodes, and omits the surrounding infrastructure. In recent years using the Power Usage Efficiency metric has been the most common way to assess the efficiency of the data centre infrastructure. The problem is that the PUE metric is not very well defined, despite the efforts of different groups. This lack of strict definition leads to problems in directly comparing the PUE of different sites, since different factors may or may not be included. There is an increasing awareness on both vendor and user sides about this problem, and this has resulted in ongoing work to create a better metric that will allow for better comparison between solutions (see sections 3.1.6 and 3.2).

Cooling has always been an important issue for supercomputing and became more important when the heat density of a cluster reached the point where using pure air cooling no longer was an economically viable option. The direct liquid cooling (DLC) approach is gaining popularity as it allows significant reduction in energy overheads. The majority of the vendors (see section 3.1) allow using warm water (W3 or W4 according to ASHRAE standards), thus enabling free cooling mode for the Data Centre cooling infrastructure throughout the year.

3.1 Cooling Systems

In the last two years one could see the return of liquid cooling in HPC. Abandoned in the 90s, liquid cooling is returning because of the increasing heat density of the servers and a rising awareness of the cost of operating the infrastructure. Below we present a few examples of how major vendors are coping with this subject. In addition we show Intel`s point of view on the future of HPC. While Intel is not a system vendor its voice is very significant as Intel-manufactured chips are dominating current HPC systems.

Using liquid as a coolant is very effective but causes lots of practical problems, including cost. Therefore, the current generation of DLC machines usually have only CPUs, GPUs and memory liquid cooled, since these are the main heat sources in the machine. The rest of the heat, which is usually 20-30% of the total energy, has to be cooled by air. The usage of rear door heat exchangers is still popular for system of moderate heat density.

3.1.1 *Bull*

Bull utilizes a straightforward direct water cooling approach in blade chassis. Cold plates are used on the CPU-boards, which make it easy to adapt the solution to new processors or accelerators. In each rack a separate chassis is used to host up to 5 compute blades, power distribution and a hydraulic module including a pump and heat exchanger. Up to 90 servers (possibly with accelerators) can be hosted in a single cabinet. The maximum power draw of this solution is 45 kW per rack.

3.1.2 *Cray*

Cray has chosen a completely different solution for water cooling. Cray does not use liquid cooling on the boards but uses air circulated by fans horizontally and sideways through the racks. The fans are located between the racks and the water-to-air heat exchangers are located in the compute racks. The air passes through all racks, cooling itself through fans and heat exchangers between each rack. Cray believes the following is true of a pure direct liquid cooling, and that its solution combines the best of liquid and air:

- Is more efficient.
- Allows higher temperature.
- Reduces chiller size or possibly removes them completely.
- Has a higher cost.
- Less flexible board design.
- Doesn't allow all components to be water-cooled.

3.1.3 *HP*

HP has designed a solution that is highly optimized for water cooling, with the goal of having 100% water cooling, a compact enclosure and easy maintenance and setup. HP has used many innovative concepts in the design such as:

- Closed loop with heat pipes on the CPU board.
- Heat is transferred from the board to the rack with metal-to-metal mechanical pressure heat transfer.
- The "dry connect" is cooled by a water circuit in the rack called "water wall". The dry connect makes it possible to replace blades without shutting the system down.
- The water wall is under negative pressure to minimize the risk of water leakage.
- An Intelligent Cooling Distribution Unit takes the space of half a rack and includes a heat exchanger between the rack water circuit and the facility water circuit plus the control system. This distribution unit can handle several racks up to 320 kW.
- The racks are connected with special piping with valves included, which is delivered in sections done in the factory.

All steps in the heat transfer limit the output temperature but the solution is compact and easy to maintain. In this solution CPU, GPU and memory are directly water cooled, the rest of the components are cooled by air. HP designed their system in a way that it uses the water from the facility loop first to cool down the air inside the racks and then the same water is used to directly cool down the hot components. This approach limits the maximum inlet temperature to 32°C but ensures that no additional cooling has to be provided by the facility infrastructure.

3.1.4 *SGI*

SGI provides many different cooling alternatives, from traditional air cooling to DLC cooling with cold plates. For HPC it seems that SGI can provide both a solution that uses air cooling of the blades but with a water-cooled water-to-air heat exchanger in what is called an M-cell. This solution is in principle the same as Cray's but with a different physical configuration with a large rack with 4 enclosures for blades and with heat exchanger and fans included.

For even higher loads, SGI can utilize cold plate technology with water cooling to the cold plate. In this case the M-cell rack can be equipped with a water-to-water heat exchanger and a pump for the water to the cold plates, making this configuration similar to the one presented by HP: one Cooling Distribution Unit that can handle several racks of compute blades.

3.1.5 *Other vendors*

DLC solutions in form similar to the ones presented above are also available in the portfolio of other system vendors (e.g. Fujitsu, Huawei, Megware, Eurotech and Iceotope). There are also solutions that allow for retrofitting of traditional, air-cooled servers with the infrastructure required to cool these directly with water (e.g. CoolIT and Asetec).

3.1.6 *INTEL - Trends and Topics in HPC Infrastructure*

Liquid cooling is a very active research topic inside Intel. Their working group is considering all types based on how close the liquid gets to the components. There can be at least 5 different types, each with its own issues and various paths of development. The development may lead to liquid cooled optimized SKUs with vertical integration. Vertical Integration means understanding system type and implications for full liquid cooling and / or immersion, based on density needs, accounting for related challenges in Watts/m^3 as much as Watts/cm^2. Each type has its own mixed temperature issue, e.g. ASHRAE W3 and W4 for compute but needs to consider cooler temperatures for comfort, dehumidification, I/O and Storage racks. The water loop needs further temperature considerations such as condensation and water temperature at the facilities / IT loop heat exchanger inlet. Finally, water quality and maintenance are the main issues. When considering water cooling, especially for warm water, all the requests should meet ASHRAE water quality guidelines and ASHRAE standard definitions.

3.1.7 *HPC and city infrastructure*

Some utility companies; like Fortum in Stockholm, Sweden; is advocating placing data centres in urban areas to be able to reuse the generated heat in their district heating networks. The Fortum utility company is able to produce more than 250 MW of cooling power, which is at least an order of magnitude more than any European HPC centre. While the cooling is provided to both commercial and private consumers, local data centres are consumers of more than 50% of the cooling capacity. The company is cooperating with the local data centres and built a custom facility design to capture and re-use heating from the local data centres. The system was designed and commissioned in 2003 so it is working with chilled water (4 / 16°C) and is able to produce 70°C water that is distributed in city heating network. A data centre facility in Stockholm was presented as an example case where the cooling is provided by city cold water distribution system and, at the same time, the computers are heating up water used for heating in Stockholm.

Fortum believes that re-using this kind of waste energy may be the cheapest source of energy and, especially taking into consideration rapid growth of the IT sector, it is wasteful not to exploit this opportunity.

3.2 Efficiency metrics

To optimize any aspects of the computing infrastructure, including the energy efficiency, one needs a well-defined set of metrics that allow for direct comparison between different solutions. The most popular PUE metric is neither defined well enough nor provides a deep insight into specifics of the machine to be treated even as a de-facto standard. It is however a starting point for future work that aims to define a better metric. The importance of having such a metric is acknowledged both by industry and academia resulting in many, sometimes very similar, approaches to the problem.

3.2.1 *Mobilizing the HPC Community for Improving Energy Efficiency: Energy Efficiency HPC Working Group*

For the third consecutive year Natalie Bates has participated in the Workshop presenting the current status and ongoing work of the Energy Efficiency HPC Working Group which she chairs. This working group provides a forum for sharing of information and taking collective action, driving energy conservation measures and energy-efficient design in HPC. The current membership, which totals more than 450 people, consists approximately of 50% government agencies (many Top100 sites), 30% vendors and 20% academia. The work is organized into three sub-groups: Infrastructure, Computing Systems, and Conferences.

The first group, Infrastructure, deals with the efficiency of HPC facilities, and has introduced new metrics to support PUE: ITPUE (captures inefficiencies in fans, liquid cooling, PSU, etc.), TUE (the product of ITPUE and PUE, which gives a more real value for the site's energy efficiency), ERE (energy reuse effectiveness), etc. This group is also studying the possibilities of demand response, whereby an HPC centre works with the electricity provider to manage their large energy requirement to maximize efficiency and grid stability.

The Computing Systems group is centred around improving power measurement methodologies, with the introduction of higher-quality power measurements (called Level 2 and Level 3) which have been included in the Green500 (at the moment submitting these values is optional). Part of their work also includes ways of linking these measurements to procurements and benchmarks to help pressure vendors.

The Conferences team organizes actions (presentations, round tables, paper submissions, etc.) during SC and ISC, as well as more topic-specific events.

3.2.2 *ITUE and TUE metrics*

Power Usage Effectiveness (PUE) metric was introduced in 2006 by Malone and Belady and it is now developed and agreed to by many institutions (EU Code of Conduct, DoE, EPA, Green Grid, ASHRAE, etc.). This metric has led Energy Efficiency drive in data centres, where in 2007 the average was about 2.7, nowadays the “best in class” big data centres are less than 1.1 or very close to it. The PUE definition is not perfect and can be improved.

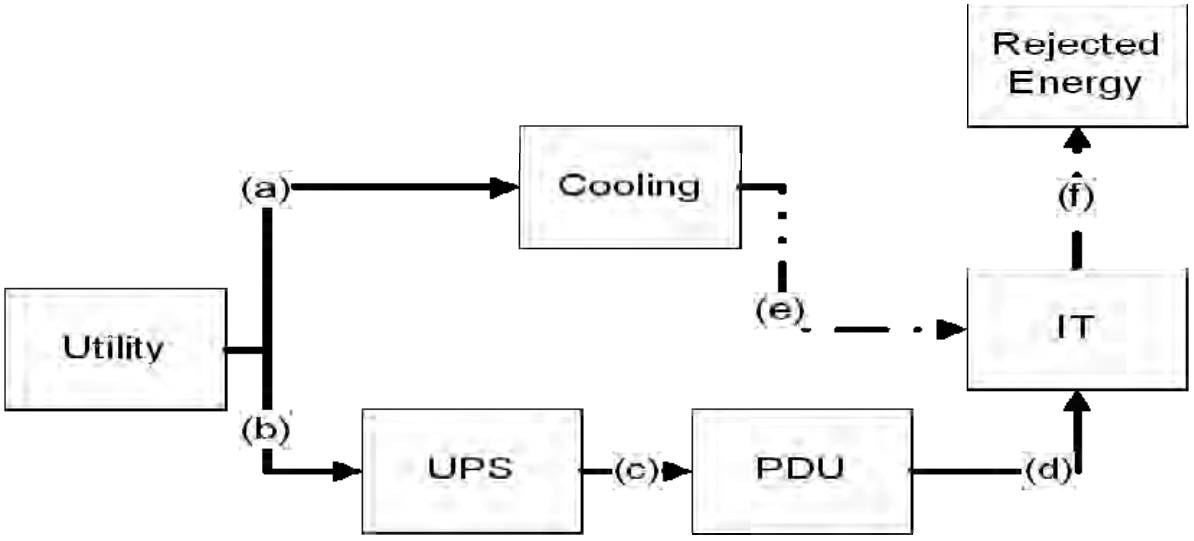


Figure 1: Diagram for PUE calculation

Referencing the diagram above, the formula to compute the PUE is:

$$PUE = \frac{Total\ Energy}{IT\ Energy} = \frac{Cooling + Power\ Distribution + Misc + IT}{IT} = \frac{a+b}{d}$$

Nevertheless the (d) input and IT definition may lead to mistakes. In fact if the IT is considered the whole "server", then the IT load will account also for fans, PSU with DC converters and other infrastructure burden inside the server, leading to a viable reduction of PUE ratio simply moving the facility burden out of the (d) term into the IT block. One way to better focus on the real IT effectiveness is to consider two different levels for PUE calculation: one outside of the server, the traditional PUE, and one inside the server itself. The latter is called ITUE and its calculation is based on the diagram unterhalb

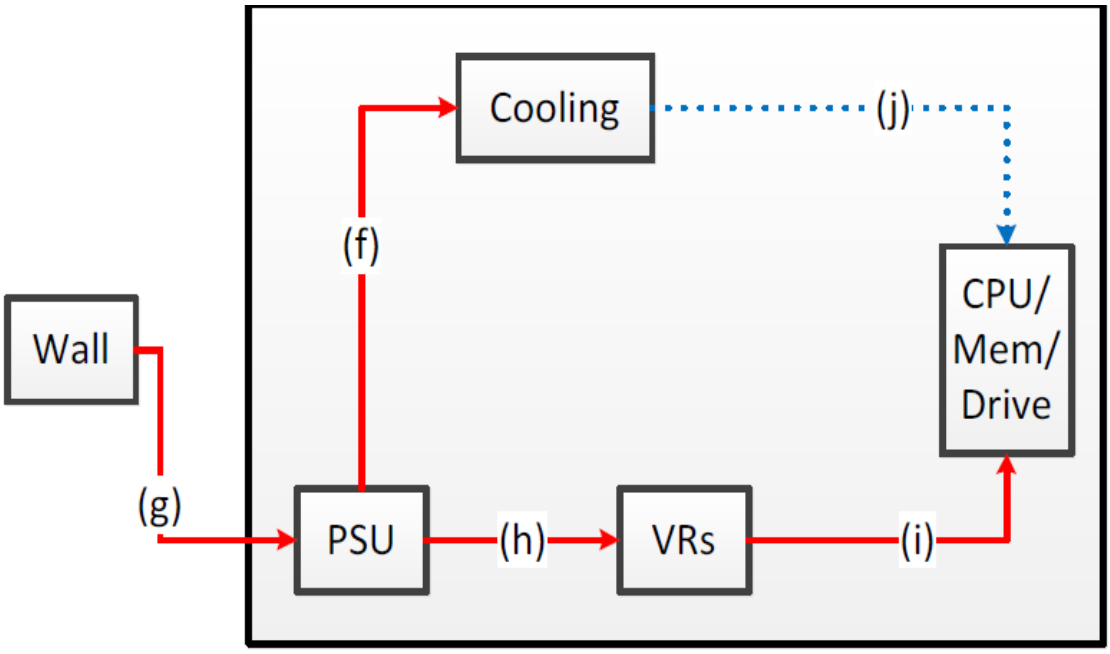


Figure 2: Diagram for ITUE calculation

and the formula:

$$ITUE = \frac{total\ energy\ into\ the\ IT\ equipment}{total\ energy\ into\ the\ compute\ components} = \frac{g}{i}$$

Chaining the two definitions leads to a new metric very focused on the IT effectiveness: TUE.

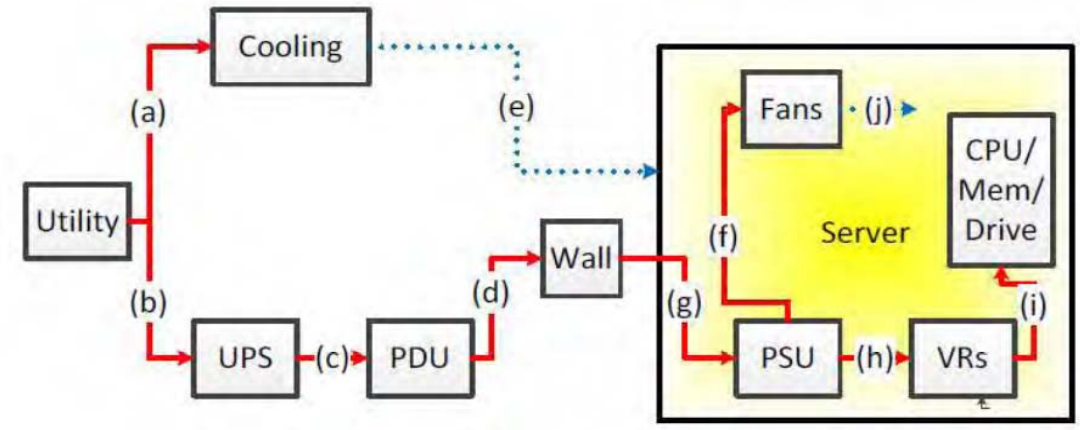


Figure 3: Diagram for TUE calculation

Considering the diagram above, TUE is defined as:

$$TUE = ITUE \times PUE = \frac{a+b}{i}$$

It can be demonstrated that better TUE means less total energy used given a fixed IT load (while this isn't always true considering only PUE).

In order to leverage the benefits of TUE, a considerable amount of work must be done to encourage industry to collect and share more and future ITUE and TUE data, and HPC is an ideal segment to develop the concepts. Over time TUE & ITUE can create the same pull and success that PUE has.

3.2.3 *Improving energy efficiency*

The real power challenge (as TUE is focusing on) is in CPU, memory and data transfer/storage efficiency. The Near Threshold Voltage (NTV) Operation efficiency has been demonstrated since 2012 and microprocessor producers are now decreasing the supply voltage at chip level. New memory technologies are promising to decrease DRAM energy footprint to a few pJ/bit while increasing the bandwidth by an order of magnitude. Finally, 3D integration of compute, I/O, and memory is the only solution for energy efficient bandwidth scaling. There can be various levels/solutions and issues in integrating the components:

- In package memory – to be closer to the CPU, stacked memory.
- Fabric integration – giving package connectivity.
- Advanced switches with higher radix and higher speeds – to get a closer integration of compute and switch.
- Silicon Photonics – lowering costs, outstanding performance (but thermal issues do exist).
- Liquid cooling for 50-100 kW racks; power supply directly with 400 V AC 3-phase or 480 V AC 3-phase.

3.3 Power and energy aware tools

The push towards increasing the scale and performance in HPC has increased energy consumption in supercomputing centres. As a result a significant amount of money and natural resources is consumed. Furthermore peak performance is rarely attained, therefore a percentage of this energy results in little or no performance gain for real life applications.
High Performance, power and energy aware computing focuses on several methods and techniques in order to reduce the amount of energy consumed in large supercomputing centres and increase the efficiency of supercomputer usage taking into account energy consumption. Such techniques include the following: improvement in cooling systems, improvement in power supply technology, development of processors that offer higher Flops/Watt, usage of power and energy aware software tools for the management of systems and user workflows and application tuning for energy efficiency.

In the following sections we focus on the last two methodologies for reducing the energy consumption, namely power and energy aware software tools and application tuning. We describe related technologies and tools as identified in our recent visit to the International Supercomputing Conference (ISC) 2014 [3].

3.3.1 *LSF*

Platform Load Sharing Facility (LSF) is a workload management platform (job scheduler) for distributed HPC environments. It can be used to execute batch jobs on networked Unix and Windows systems on many different architectures.

Energy aware scheduling for IBMs Platform LSF is based mainly on the following techniques:

a) Scheduling based on environmental factors, such as machine temperature.
b) Power on/off idle nodes based on the observed workload in the supercomputing system.
c) Reduction of power consumption of idle nodes.
d) Optimization of power consumption of nodes that are running jobs based on the actual workload.
e) Detailed accounting of energy consumption per user job.

Techniques b, c and d were introduced with LSF 9.1.2, released in December 2013, and will be described here in more detail as they are the state of the art tools for achieving more efficient usage of the systems in terms of performance and energy consumption. Technique b, i.e. powering on/off nodes works quite well at small scale, however in modern supercomputing centres with thousands of nodes this is not always optimal and can lead to several problems and delays.

Traditionally all compute nodes of an HPC system run at the nominal frequency by default. The active nodes run all the jobs at the default speed at all times and idle nodes consume a lot of energy since there isn't any special setting for the power consumption of idle nodes. LSF provides the functionality of setting the "ondemand" CPUFreq governor of the nodes and offers the ability to intelligently put idle nodes into S3 state based on predefined policies.

Host Power Management in LSF is handled in 4 different ways:

- Via manual management scripts such as "badmin".
- Via policy-driven power saving based on the time windows (i.e. from 19:00 to 08:00 put idle nodes to power saving mode S3).
- Via power saving aware scheduling by scheduling jobs to use idle but not power saved nodes first, and then wake up nodes on demand when new jobs are to be executed by the batch system.
- Via customizable power control scripts i.e. OOB xCAT

Active Node Power Consumption Optimization is based on the ability to set a specified frequency on CPU/node level for a given job or application or queue and to intelligently tag the job for selecting the optimal CPU frequency based on energy and performance characterization of the job or application. To achieve this, LSF evaluates and calculates the power profile of the system's nodes and saves them in a database. The system administrator sets a default frequency for the whole set of nodes of the HPC system. When a new job appears for the first time in the system, or a user schedules an application, the user can tag it with his preferred name, and run it with the default frequency. LSF, using the hardware provided system performance and power consumption metrics, collects energy consumption information and generates a profile for the application that is then saved in the database. When a user re-submits the same application, he/she can specify an associated energy policy; either minimize time to solution or minimize energy to solution; so that LSF can select the optimal CPU frequency for the given application and energy policy.

Minimize time to solution policy aims to provide better application performance by running the application in a higher frequency than the default one. The administrator sets the

minimum performance improvement expected from an application running at a higher frequency than the default. Based on previous runs of the same applications, LSF selects the optimum frequency of the hosts CPU to run the applications.

Minimize energy to solution aims to save energy by setting a maximum allowed performance degradation for applications when lowering CPU frequency. The maximum allowed performance degradation is set by the administrator, and based on previous runs of the same application LSF selects the lowest frequency that meets the set requirement.

Among PRACE partners, LRZ has experimented with and is working with these capabilities. They have demonstrated that selecting the optimal (in terms of energy efficiency) frequency for an application may lead to a substantial reduction of energy usage.

LRZ evaluated the predictions of runtime and power consumption that IBM LoadLeveler (another scheduler, which has functionality that is being transferred to LSF) performs for applications running at a given CPU frequency and then compared predictions with the actual measurements. The experimentation was performed using several applications as well as synthetic benchmarks. The results have shown that the predictions have high accuracy therefore the technique could be used with the real application portfolio.

Energy to Solution and Energy Delay Product were the metrics that have been evaluated, in order to draw conclusions about the best policies to use when deciding the most appropriate frequency that any given user application should run at.

Applications should be offered a higher than default frequency if the runtime decreases by specific percentages when this frequency is provided. These policies, apart from providing an efficient utilization of the system, also provide an incentive to users to optimize their codes and tune their applications.

3.3.2 *Adaptive Computing (MOAB)*

Adaptive Computing, at ISC'14 (June 2014), announced its Moab HPC Suite-Enterprise Edition 8.0 (Moab 8.0) that provides Advanced Power Management features. The new features include:

- Clock frequency control
- Additional power state options

With clock frequency control, administrators can adjust CPU speeds to align with workload processing speeds through job templates. In addition, administrators can manage multiple power states and automatically place compute nodes in new low-power or no-power states (suspend, hibernation and shutdown modes) when nodes are idle.

The features seem to be similar to the ones provided by IBM's platform LSF.

3.3.3 *Allinea Tools*

Allinea provides tools to improve the efficiency of HPC system usage, by reducing development time and increasing application performance by offering tools designed to handle demanding applications and environments. During ISC'14 Allinea announced new features in their products that introduce power and energy efficiency from the application point of view.

The main Allinea products that support the lifecycle of application development are Allinea DDT (debugger) and Allinea MAP (profiler).

The new features in Allinea MAP version 4.3 include the integration of energy metrics that allow the application developers to identify lines in their code that are considered consuming

hotspots therefore enabling the user to optimise them for energy consumption and thus minimizing energy footprint. Further to that Allinea MAP adds accelerator metrics to identify the impact of accelerators in energy consumption of applications.

With the above functionality application developers and/or operators who have access to the source code of the applications can identify:

- if the application is running in an efficient way
- if there are any software or hardware bottlenecks that affect performance
- if the combination of application parameters/libraries is optimal
- the scale that should be used for a job in order to maximize efficiency
- which functions are using the most energy
- what points in time the app uses the most energy
- impact of co-processors to energy consumption

Metrics that Allinea MAP can collect or calculate from hardware include: total energy, peak power, average power, total system energy, total system peak power and total system average power. The tools support both multi-threaded and hybrid codes.

3.3.4 *Bull and HDEEM*

In 2013, Bull partnered with the Technical University of Dresden in a project named HDEEM: High-Definition Energy-Efficiency Monitoring. The goal is to expose very accurate power measurements (500 samples per second) to the end-user and to link it with the source code. Bull is instrumenting its nodes by incorporating an FPGA to monitor the energy consumption at relatively high frequency (500 Hz) at an accuracy level of 2%. Furthermore, Bull developed a low-level API (metrics, levers). Dresden developed a high-level API, which can be used by tools like Vampir, a well-known performance analysis tool. The data presented by Vampir can be used to identify issues like computational and communication bottlenecks, load imbalances and inefficient CPU utilisation. The purpose of HDEEM is to additionally provide precise metrics not just in performance but also in terms of power consumption. Through the use of HDEEFM different execution policies can be defined: time to result (as fast as possible), Joules to result (minimal power consumption), price to result (cheapest execution), eco impact to result (green energy). At ISC'14, Bull presented the FPGAs as an optional component in their upcoming B720 blades (B700 series blade for Intel Haswell, using Direct Liquid Cooling).

3.4 Chapter Summary

Energy consumption and its optimization has become one of the key challenges that define the design of future (also exascale) machines. One can see that the evolution of computing systems is progressing not only with pure computing power in mind but also taking care of other factors like the ability to cool down and manage the power consumption of the future systems on all levels of operation: starting from the data centre integration and ending at the level of how the silicon is integrated. In parallel, energy efficiency aspects are taken into consideration not only in design of the hardware but also in the system software and resource managers. There are also some projects that aim to create energy consumption awareness inside applications, but these are at very early stages of development.

4 Assessment of Petascale Systems

Since 2008 this work package and the similar work packages in the previous projects have been following the introduction of petascale systems in the Top500 list, from Roadrunner (the first petascale system, now decommissioned) to the current 55 supercomputers around the globe that have reached a peak performance of 1 PFlop/s or more 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 in terms of companies and roadmaps, 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 55 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 2014 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
2. The web site of the computing centre hosting a supercomputer
3. Vendor specific web sites
4. Funding agencies web sites

Detailed lists of such sources broken down in categories as specified above are available in the previous deliverables of the activity such as D5.1 [1] and D5.2 [2].

To facilitate a more efficient search among the results of Google searches we also used the HPC Market Watch Google Custom Search engine (CSE) that was specifically created within WP5 for this activity. 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.

4.1.2 *Snapshot*

On June 22nd 2014, the 43rd Top500 list of the world's most powerful supercomputers was presented at ISC13 in Leipzig, and it included 55 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.

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, 21 systems)
- Lightweight: use many low-power RISC processors (in green, 11 systems)
- Traditional: use only standard high-performance processors (in blue, 23 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 BlueGene/Q (IBM PowerPC)	17.17 / **20.13**
K Computer	RIKEN (Japan)	Fujitsu Cluster (Fujitsu SPARC64)	10.51 / **11.28**
Mira	ANL (USA)	IBM BlueGene/Q (IBM PowerPC)	8.59 / **10.07**
Piz Daint	CSCS (Switzerland)	Cray XC30 (Intel Xeon / NVIDIA Tesla)	6.27 / **7.79**
Stampede	TACC (USA)	Dell PowerEdge (Intel Xeon / Intel Xeon Phi)	5.17 / **8.52**
JUQUEEN	FZJ (Germany)	IBM BlueGene/Q (IBM PowerPC)	5.01 / **5.87**
Vulcan	LLNL (USA)	IBM BlueGene/Q (IBM PowerPC)	4.29 / **5.03**
Anonymous	Government (USA)	Cray XC30 (Intel Xeon)	3.14 / **4.88**
HPC2	Eni S.p.A. (Italy)	IBM iDataPlex (Intel Xeon / NVIDIA Tesla)	3.00 / **4.00**
SuperMUC	LRZ (Germany)	IBM iDataPlex (Intel Xeon)	2.90 / **3.19**
TSUBAME 2.5	GSIC (Japan)	NEC/HP ProLiant (Intel Xeon / NVIDIA Tesla)	2.79 / **5.74**
Tianhe-1A	NSCT (China)	NUDT YH MPP (Intel Xeon / NVIDIA Tesla)	2.57 / **4.70**
Cascade	EMSL (USA)	Atipa Visione (Intel Xeon / Intel Xeon Phi)	2.54 / **3.39**
Pangea	Total EP (France)	SGI ICE X (Intel Xeon)	2.10 / **2.30**
Fermi	CINECA (Italy)	IBM BlueGene/Q (IBM PowerPC)	1.79 / **2.10**
Edison	NERSC (USA)	Cray XC30 (Intel Xeon)	1.65 / **2.57**
Anonymous	ECMWF (UK)	Cray XC30 (Intel Xeon)	1.55 / **1.80**
Anonymous	ECMWF (UK)	Cray XC30 (Intel Xeon)	1.55 / **1.80**
Pleiades	NASA (USA)	SGI ICE X (Intel Xeon)	1.54 / **2.11**
DARPA TS	IBM DE (USA)	IBM Power 775 (IBM POWER7)	1.52 / **1.94**
Blue Joule	STFC (UK)	IBM BlueGene/Q (IBM PowerPC)	1.43 / **1.68**
Spirit	AFRL (USA)	SGI ICE X (Intel Xeon)	1.42 / **1.53**
Archer	EPSRC (UK)	Cray XC30 (Intel Xeon)	1.37 / **1.65**
Curie TN	TGCC (France)	Bull B510 (Intel Xeon)	1.36 / **1.67**
Hydra TN	RZG-MPG (Germany)	IBM iDataPlex (Intel Xeon)	1.28 / **1.46**
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)
Helios	IFERC (Japan)	Bull B510 (Intel Xeon)	1.24 / **1.52**
Garnet	ERDC (USA)	Cray XE6 (AMD Opteron)	1.17 / **1.51**
Cielo	LANL (USA)	Cray XE6 (AMD Opteron)	1.11 / **1.37**
DiRAC	EPCC (UK)	IBM BlueGene/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**
Quartetto	RIIT-KU (Japan)	Hitachi/Fujitsu PRIMERGY (Intel Xeon / NVIDIA Tesla / Intel Xeon Phi)	1.02 / **1.50**
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**
Lomonosov	RCC (Russia)	T-Platforms T-Blade (Intel Xeon / NVIDIA Tesla)	0.90 / **1.70**
Anonymous	RPI (USA)	IBM BlueGene/Q (IBM PowerPC)	0.89 / **1.05**
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**
COMA	CCS-UT (Japan)	Cray CS300 (Intel Xeon / Intel Xeon Phi)	0.75 / **1.11**
Hydra AN	RZG-MPG (Germany)	IBM iDataPlex (Intel Xeon / NVIDIA Tesla)	0.71 / **1.01**
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**
Anonymous	Unknown (USA)	HP ProLiant (Intel Xeon)	0.50 / **1.13**
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)	0.43 / **1.09**
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 merging the resulting conclusions, the market can be described in much more detail, providing a better understanding of the underlying environment. When a percentage number is given in this section it refers to the share of the total number of petascale systems, unless stated otherwise.

4.1.3.1. *Year of construction*

Peak performance of 1 PFlop/s was reached for the first time in 2008 (in publicly listed computers), yet the oldest petascale systems in production today are all from 2010. In fact, almost three fourths of the systems in the market watch have been built or updated since 2012, which is still the best year for petascale system introductions. There does seem to be a slowdown in the last two years (it is still early to predict the total number of petascale systems that will be introduced in 2014, but even the highest estimates only equal the results of 2012), possibly due to the long-lasting global economic slowdown.

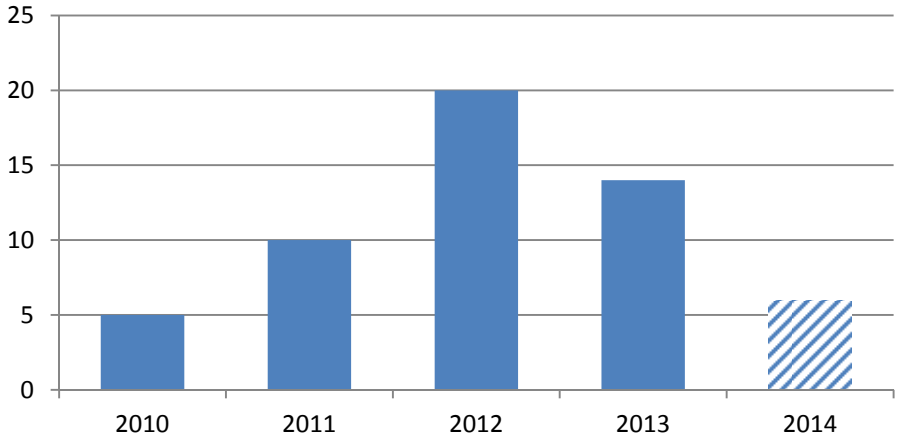


Figure 4: Petascale systems by year of deployment

4.1.3.2. *Country*

China has been leading the Top500 list for the past year with their Tianhe-2 system, but the USA still has significantly more petascale sytems (23, almost four times more than China) and therefore dominates the market (42%). Country-wise, China and Japan tie for second place with 6 each (11% share), slightly ahead of Germany and the UK with 5 each (9% market share). France and Italy follow them with 3 and 2 systems respectively. Spain, Russia, Australia, Saudi Arabia, and Switzerland each have one. The EU as a whole contains 30% of all petascale systems, which would be considered second place between the USA and China.

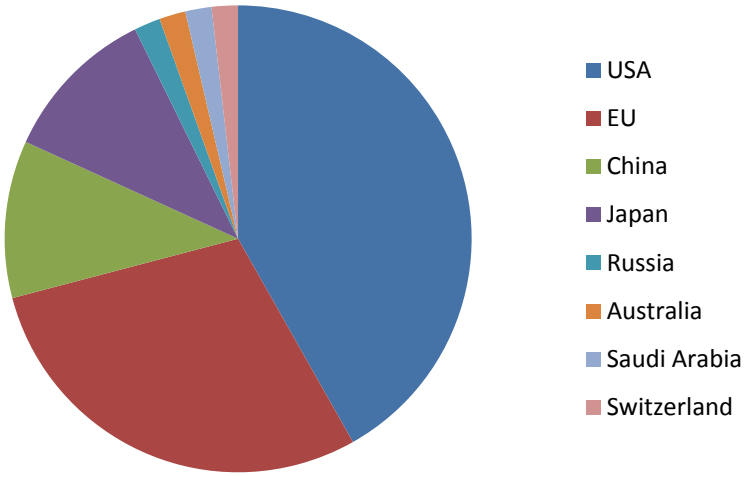


Figure 5: Petascale systems by country

Architecture-wise it is interesting to see that 5 out of 6 of China's petascale computers are accelerated (83%), compared with 50% for Japanese and only 35% for American systems. In contrast, in the EU only 2 systems use accelerators to achieve their petascale performance (12.5%), preferring traditional clusters (62.5%) and lightweight MPP systems (25%).

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 4.3 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 an order of magnitude above the minimum).

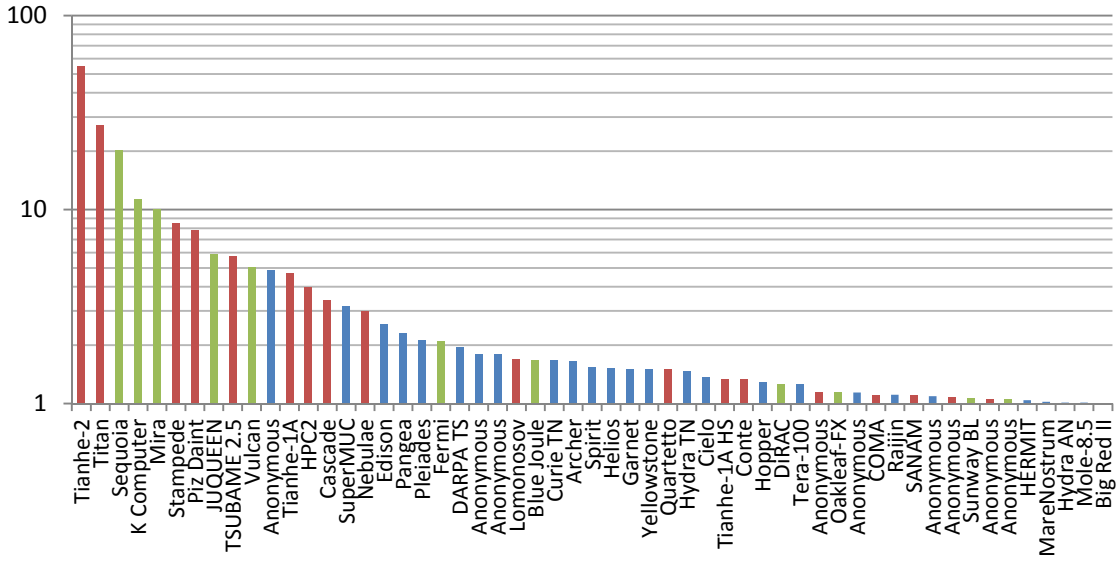


Figure 6: Peak performance of petascale systems (in PFlop/s)
(Red = accelerated, green = lightweight, blue = traditional)

It is worth noting that only one of the top 15 systems in terms of peak performance is based on a traditional architecture, even though these systems are the most popular configuration for petascale.

4.1.3.4. *LINPACK performance*

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 PFlop/s but the median is only 1.3 PFlop/s).

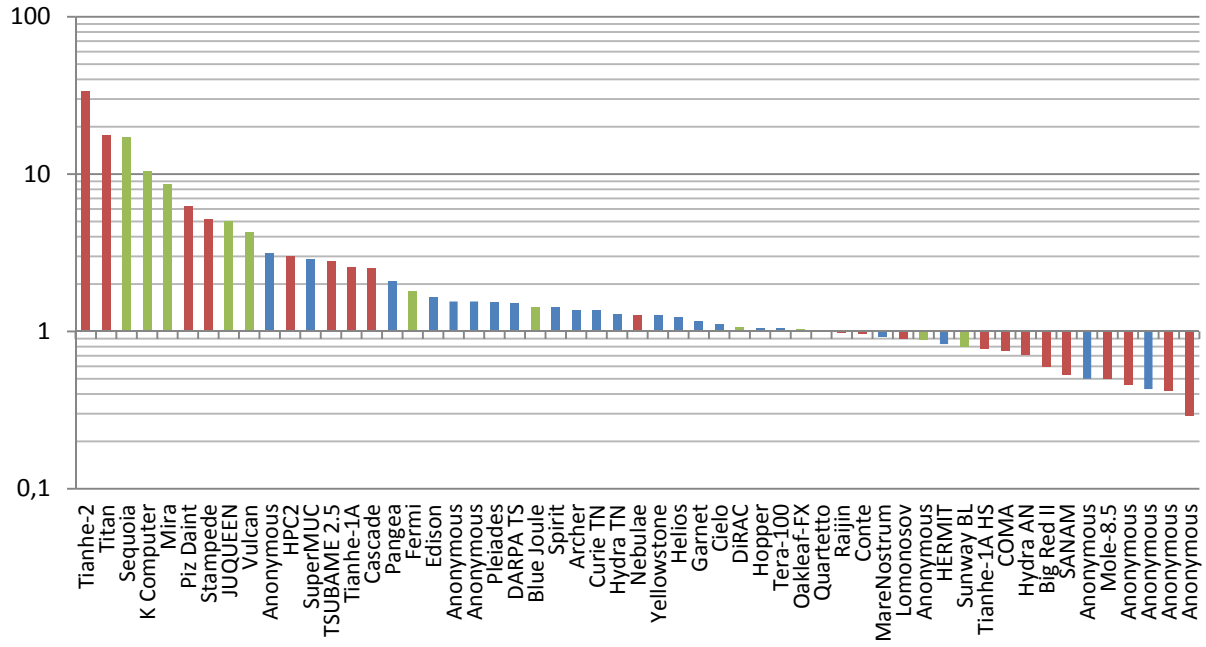


Figure 7: LINPACK performance of petascale systems (in PFlop/s)
(Red = accelerated, green = lightweight, blue = traditional)

With regards to architecture, traditional systems tend to occupy the area around the median performance, while accelerated and lightweight architectures dominate the highest performers. On the low end there is a clear majority of accelerated systems, which usually have worse LINPACK efficiency.

LINPACK scores do not fully reflect all aspects of performance since they represent only one benchmark with a very specific type of computation. This is a topic that has been in discussion for some time, and there is a section in the following Business Analysis chapter detailing new proposals for benchmarks.

4.1.3.5. *Vendor*

IBM and Cray together dominate more than half the market, with IBM slightly ahead (27% vs. 24%). HP has 6 full systems and one joint venture with NEC (TSUBAME 2.5), giving them around 12% market share (in contrast to their leading position in the entirety of the Top500 list). 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 petascale systems are all from different vendors: Dell, Dawning, T-Platforms, Tyan/IPE, NRCPCET, Atipa, and Adtech.

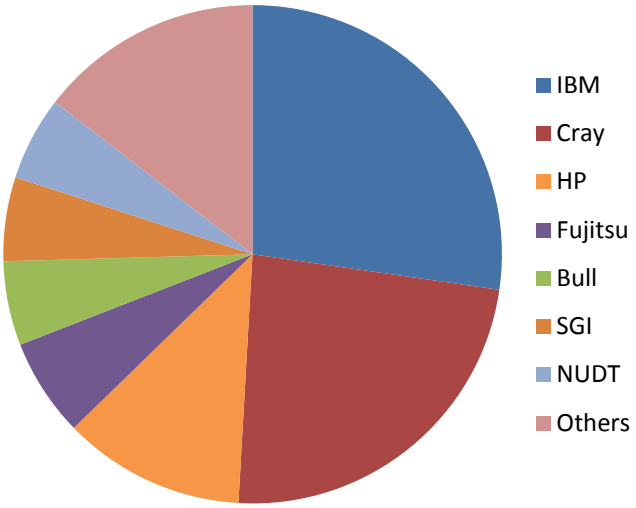


Figure 8: Petascale systems by vendor

Around half of IBM systems are Blue Gene/Q, with a lightweight architecture. The rest are mostly traditional (33%), with only 2 IBM accelerated systems. Cray, on the other hand, prefers the traditional (69%) and accelerated (31%) architectures, with absolutely no models based on a lightweight architecture. HP deals mainly with accelerated systems (67%), with only two traditional architectures and no lightweight systems. It's telling that each of the top three vendors is focused on a different architecture, although all of them also have at least another architecture type available.

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 67% 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's PowerPC processor (available only on Blue Gene/Q systems) taking almost 15% market share while AMD's Opteron only manages 11%. The remainder of the market consists of the Fujitsu SPARC64, found in 2 systems, and the ShenWei SW1600 and IBM POWER7, each in one.

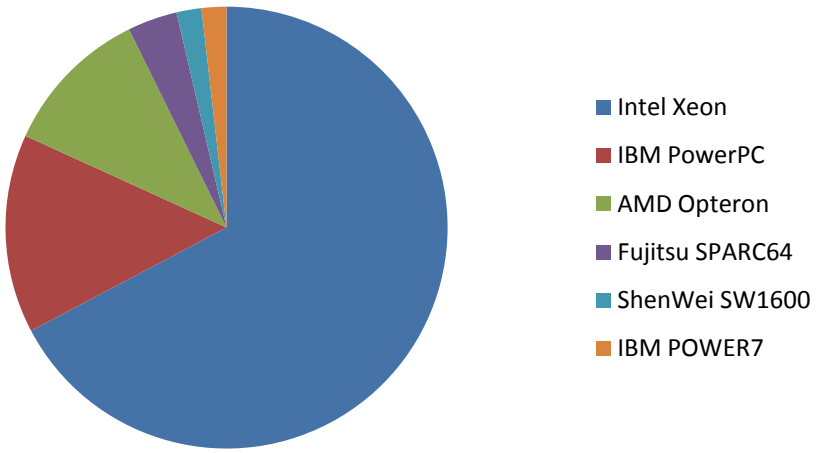


Figure 9: Petascale systems by processor

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 2.4 GHz. Traditional and accelerated

systems tend to have processor speeds around 2.6 GHz (since the accelerators are added to traditional high-performance CPUs), while lightweight architectures use CPUs that run around 1 GHz slower.

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%). Of the 21 systems that do have an accelerator, 15 use some form of NVIDIA Tesla GPGPU (71%). The next most popular accelerator, present in 5 systems, is the Intel Xeon Phi coprocessor (24%). 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.

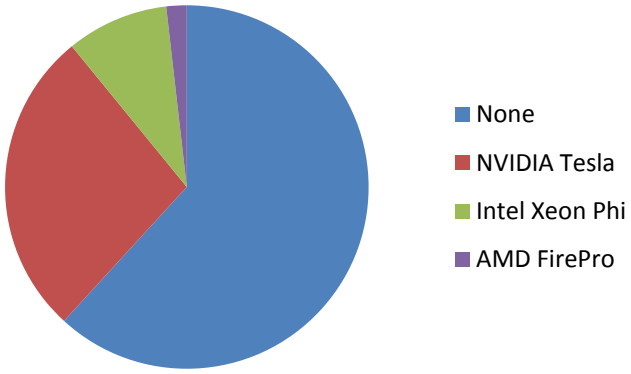


Figure 10: Petascale systems by accelerator

4.1.3.8. *CPU cores*

Core count (not counting accelerator cores) ranges from around 2,500 cores in the case of Mole 8.5, to more than 1.5M in Sequoia. The large discrepancy in the number of cores in these two systems, although partly due to their large 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 (Mole 8.5) or low-power many-core processors (Sequoia).

Analysing each of the architectures separately, and considering MFlop/s per core to level the playing field, we can clearly see the difference between the three with respect to CPU cores. Traditional systems have a performance per core of around 20 MFlop/s, while lightweight architectures perform only 13 MFlop/s per core. Accelerated systems, on the other hand, delegate much of the computation to their accelerators, allowing a performance per CPU core of around 125 MFlop/s (quite variable depending on the amount of coprocessors). Adding the accelerator cores to the CPU cores would obviously reduce this number, but it is not completely legitimate since accelerator cores are not equivalent to CPU cores (they are in fact not even equal between different coprocessors).

Looking at the chart one can see several traditional systems at the low-end, interestingly all based on Opteron processors except Tera-100 which uses older Intel Nehalem chips. This is probably due to two factors: AMD is at least one step behind Intel in terms of semiconductor process (this is also what is affecting the Nehalem chip), and the new AMD Bulldozer microarchitecture shares FP units between cores, making each less productive.

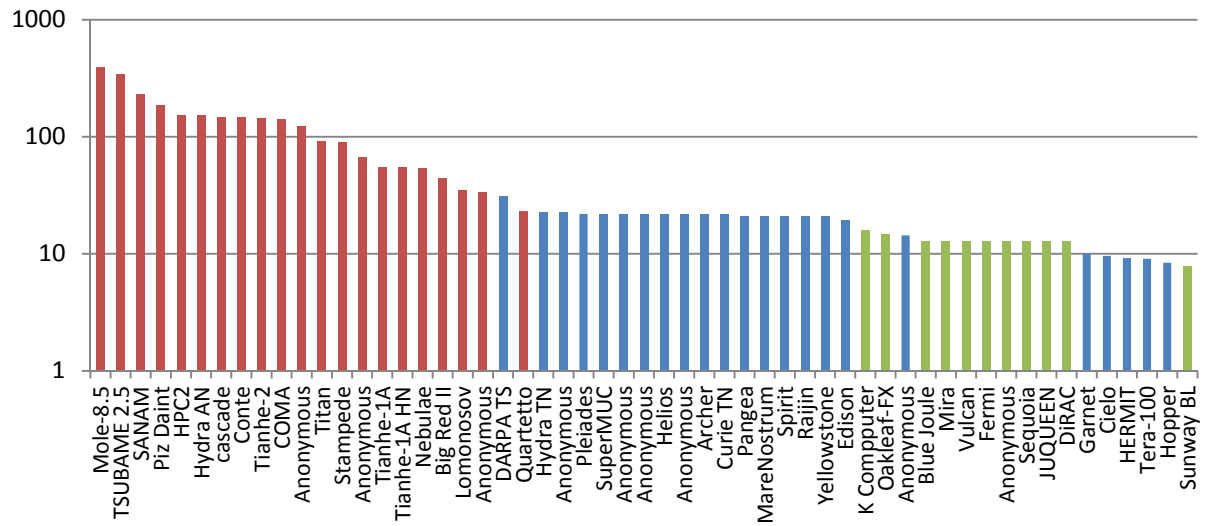


Figure 11: MFlop/s per CPU core
(Red = accelerated, green = lightweight, blue = traditional)

4.1.3.9. *Interconnects*

A total of eleven different interconnect technologies are used throughout the petascale systems, of which the most common are:

- **InfiniBand QDR / FDR** – These two interconnects represent the current industry standards defined by the InfiniBand Trade Association. 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 / Aries** – Both of these technologies were originally developed by Cray and then bought by Intel. 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. Aries removes the dependency on Opteron HyperTransport and instead uses the standard PCIe bus, as well as a new "Dragonfly" topology.
- **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.

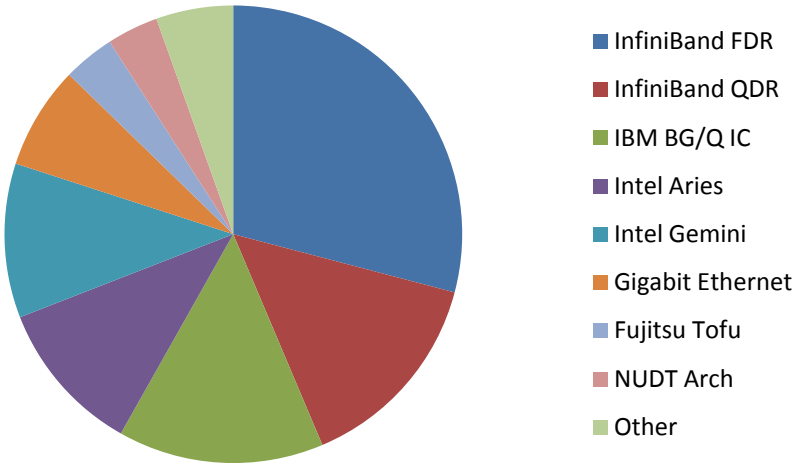


Figure 12: Petascale systems by interconnect

The most popular of these interconnects is the newest InfiniBand standard, FDR, with 29% market share (double that of the next most common interconnects). IBM's BG/Q interconnect is used solely on Blue Gene/Q machines, yet still takes almost 15% of the market, the same as the slower QDR InfiniBand. The Cray-developed and Intel-owned Gemini and Aries networks each have an 11% share, indicating that the market is in the middle of the transition from the older to the newer technology. Gigabit Ethernet is the next big player, controlling more than 7% of the total. Fujitsu Tofu and NUDT Arch each have two systems, representing a 3.6% share. The other interconnects used are the IBM P7 IC (used in the IBM DARPA prototype), 10G Ethernet (the successor of Gigabit Ethernet), and TH Express-2, the custom interconnect designed by NUDT specifically for Tianhe-2 (could be considered the successor of Arch).

4.1.3.10. *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%.

Similarly to core count, computing efficiency is very different depending on the architecture of the system. Accelerated systems average less than 60% efficiency, while lightweight and traditional set-ups have much higher efficiencies, with lightweight slightly ahead (85% efficiency on average, compared to 80% for traditional machines).

It is important to note that this ratio is strongly related to the code used to measure "real-world" performance. A system with a low efficiency running LINPACK could in theory have 99% efficiency on other benchmarks or on actual applications running on the machine. For example, the two traditional-architecture systems with efficiencies lower than 50% are both owned by private companies that probably run very specific applications for which their system has been optimized instead of LINPACK. There is more on this subject in the following Beyond HPL chapter, where new benchmarks are considered.

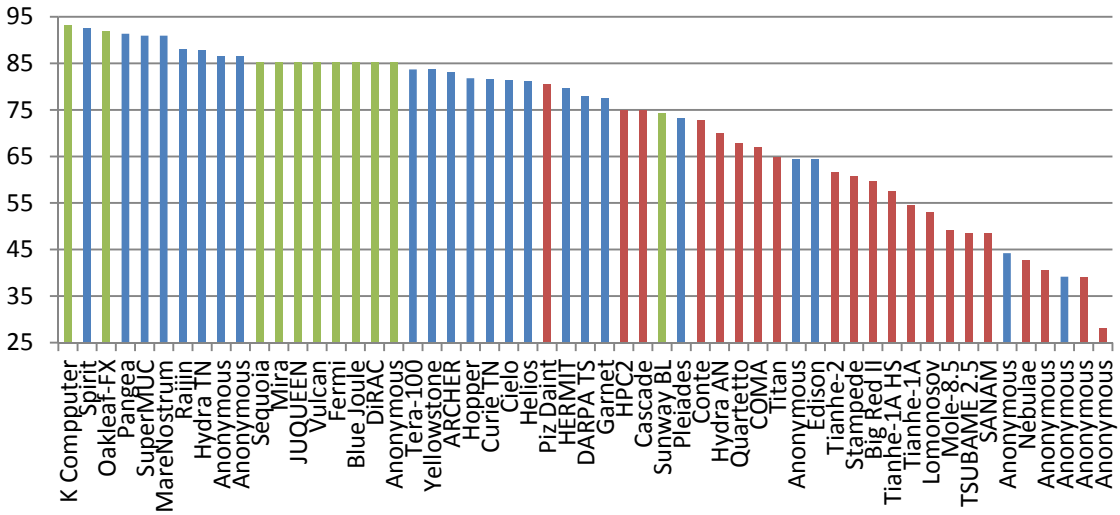


Figure 13: Computing efficiency of petascale systems (in %)
(Red = accelerated, green = lightweight, blue = traditional)

4.1.3.11. *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.

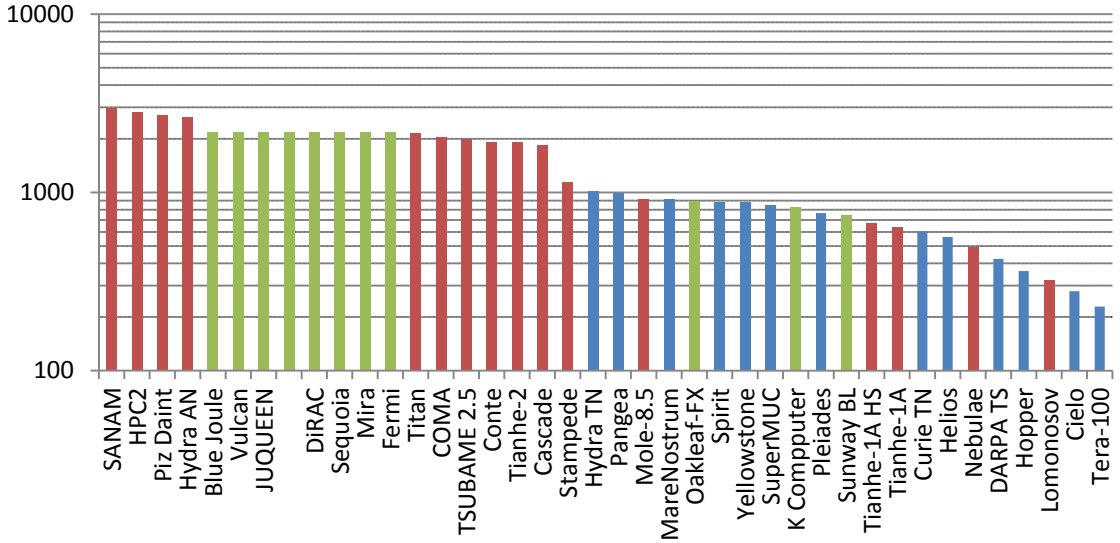


Figure 14: Power efficiency of petascale systems (in MFlop/s/W)
(Red = accelerated, green = lightweight, blue = traditional)

It is quite apparent by observing the chart that traditional systems have a hard time competing in this front with accelerated and lightweight architectures (this is not surprising if you take into account that reducing power consumption was one of the major drivers in the introduction of these new architectures). SANAM, and accelerated system based on AMD FirePro S10000 GPUs, has the highest efficiency on the list (close to 3 GFlop/s/W), followed by systems using NVIDIA K20x (the newest Tesla accelerators) and the IBM BlueGene/Q systems.

The overall average for all petascale systems is 1.4 GFlop/s/W, which can be decomposed by architecture as: 1.7 GFlop/s/W for accelerated systems, 1.8 GFlop/s/W for many-core systems, and 673 MFlop/s/W for traditional systems (considerably less than the other two).

This shows that the newer accelerated and lightweight architectures are, in general, much more power efficient than the traditional systems based exclusively on standard high-performance processors. It should be emphasized that the Green500 is based only on measurements taken while running the LINPACK benchmark, so other workloads may yield different results. Analysing the energy to solution for a representative application workload is needed when procuring a new system.

4.1.4 *Dynamic Analysis*

Having an overview of the current situation of the world-class HPC market is useful, but it is much more interesting to have a 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.

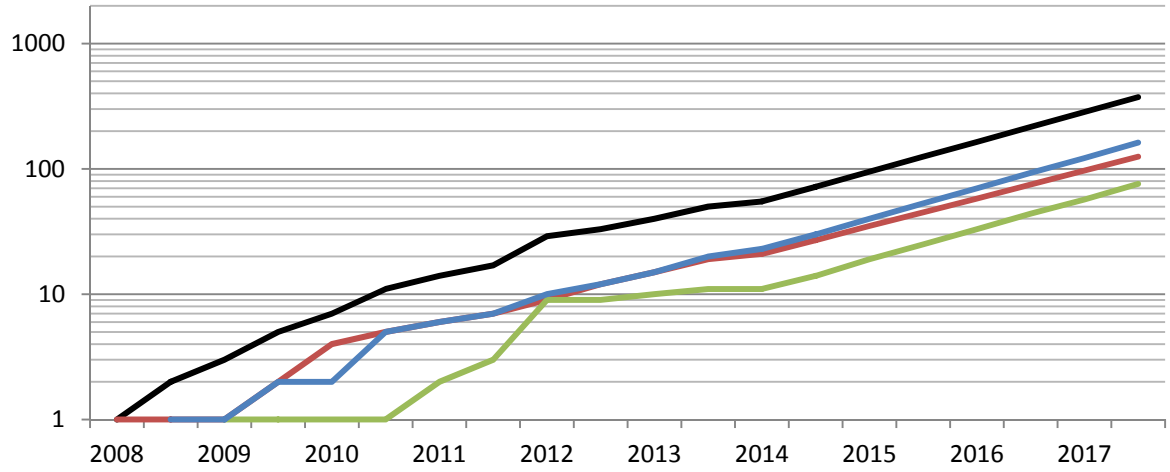


Figure 15: Evolution and prediction (from 2014 onwards) of the number of petascale systems Total (in black), broken down by architecture: accelerated (in red), lightweight (in green), and traditional (in blue)

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. It seems that the boom in accelerated and lightweight architectures seen in the first years of petascale (2008-2012) is now subduing, with traditional systems now slightly in the lead. The case of lightweight systems is especially notable, where there was a large increase in the 2010-2012 timeframe and very little growth since then.

4.1.4.2. *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 led 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.

The United States remains the clear head of the group, maintaining around 40% market share throughout the years despite the introduction of new competitors (mainly China).

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 repeated in 2013 (and have maintained until now). However, there appears to be a slight slowdown in the number of high-end systems coming from China in the last year, allowing Japan to finally catch up.

Japan entered the petascale race two years later but has introduced a somewhat continuous stream of petascale systems to the point where they are now tied with China in terms of market share. They also had one of their systems in the top position for the entire year 2011, joining USA and China as the only 3 countries to have achieved this with a petascale system.

Germany was the second country to have a petascale system on the Top500 list in 2009, and is still one of the main competitors right behind China and Japan. The UK, which entered much more recently (2012), has now surged to tie with Germany as petascale leaders in Europe, followed by France, Italy, Spain, and Switzerland. Together the EU nations make up one fourth of the market share, which would in fact be second place between China/Japan and the USA.

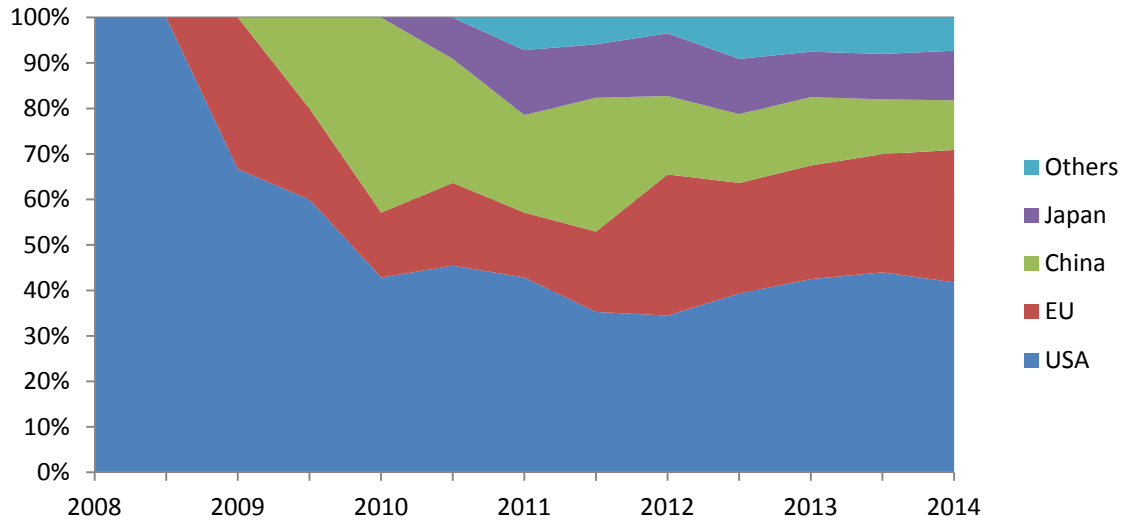


Figure 16: Evolution of the country of petascale systems

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

4.1.4.3. *Performance*

Performance of the top system has been more or less doubling each year (both in terms of theoretical peak and LINPACK score). If this trend continues, 100 PFlop/s systems should be available next year, in 2015, and the first exascale machine (in peak performance) will appear sometime in 2018.

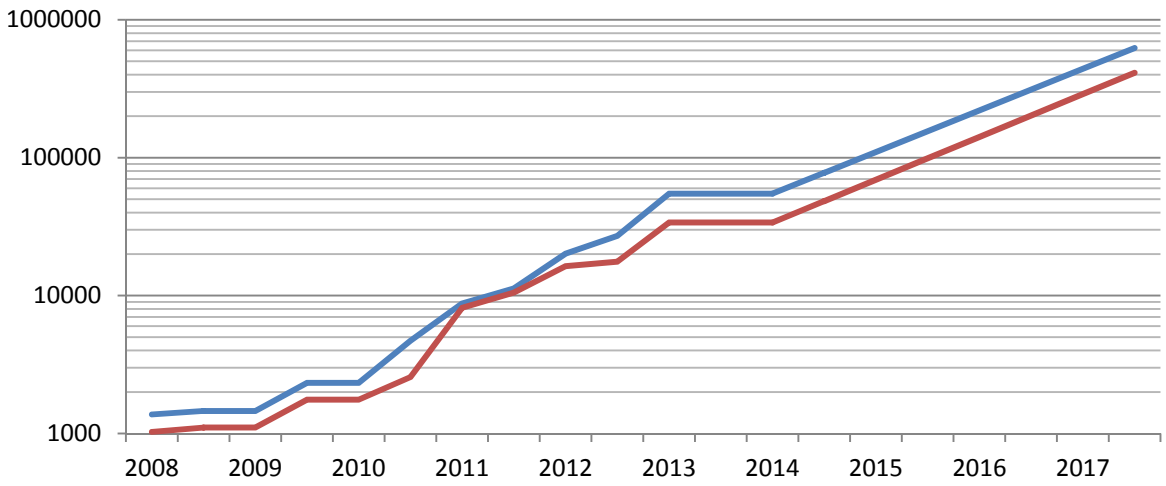


Figure 17: Evolution of maximum LINPACK (red) and peak (blue) performance (with predictions starting from 2014)

4.1.4.4. *Vendor*

Petascale computing arrived thanks to the two best known HPC vendors in history: Cray and IBM, which have managed to stay in a dominant position and together have held around 50% of the market throughout the years.

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, XK7, and now XC30.

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 2013, 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 high-end systems, but instead prefers to join in when the market is more open and lucrative. As petascale becomes more and more mainstream, HP will surely grow their market share.

SGI, Bull and Fujitsu have been the first second-tier vendors to enter the petascale market in 2010 and 2011, and have introduced new systems every year since then. The problem is that their volume is not enough to compete with the "big 3", so their market share is at best constant.

Since NUDT is not a commercial vendor but an experimental institution, it is logical that they are not striving to add many systems or grow their market share, but instead have been releasing a very high-end computer about every 2 years.

Some smaller vendors are starting to enter the petascale business (Adtech, Atipa, T-Platforms, …), limiting the growth potential of the main players. It is impossible to tell how many of these new competitors expect to grow and possibly challenge HP, SGI, Bull, and Fujitsu (or even Cray and IBM), but the heterogeneity of the petascale landscape is creating opportunities for smaller companies to flourish while the big enterprises try to maintain their ground.

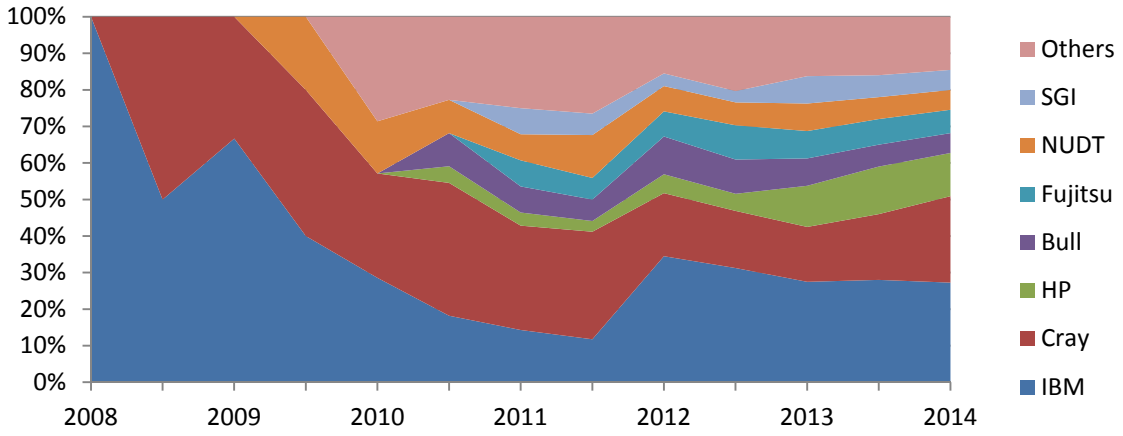


Figure 18: Evolution of vendors of petascale systems

4.1.4.5. *Processor*

It is interesting to see in the distribution of processors that Intel, the overwhelmingly dominant manufacturer of processors for both consumer computers and HPC, 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 70% of the market to themselves.

AMD, which usually tries to take a part of Intel's majority share, has in this case started with the dominant position and has been steadily losing ground since then. This coincides with a change in long-term objectives for AMD, where the server market drops in importance to strengthen their position in the mobile and mainstream sectors.

With the introduction of their PowerPC-based Blue Gene/Q in 2012 IBM jumped 20% in market share (mostly lost by AMD and, slightly less, Intel), and has been more effective at maintaining their position than AMD. IBM also has a POWER7-based petascale system, which might help them add a little more to their market share in the future. In any case, the sale of their server division to Lenovo could effectively end their HPC involvement (more on this in the Business Analysis chapter).

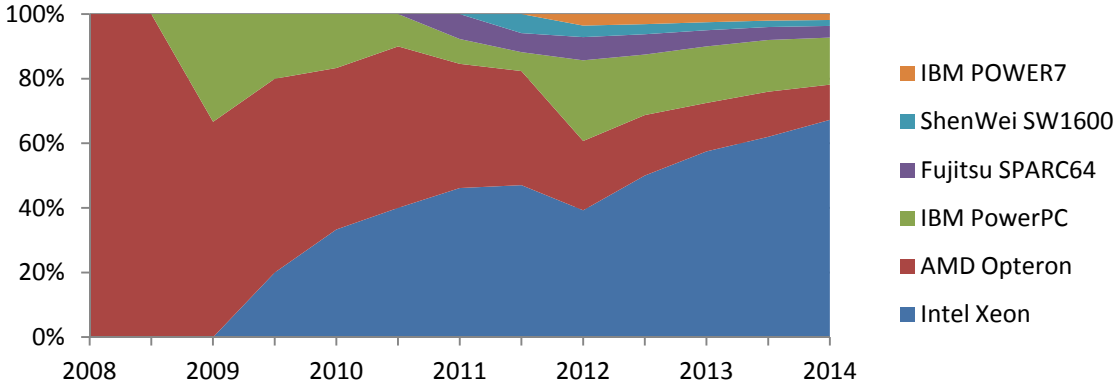


Figure 19: Evolution of processors used in petascale systems

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 3 years with respect to these two processors, so it is hard to determine whether they will return in updated versions.

4.1.4.6. *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 40% accelerator usage, so it is not clear whether accelerated petascale systems will ever be the norm, although it is also clear that they will be an important part of the mix.

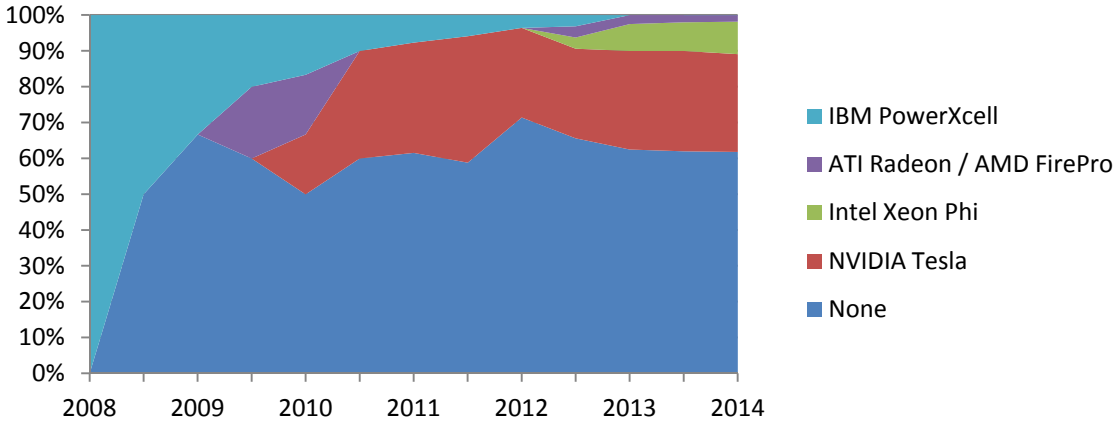


Figure 20: Evolution of accelerators used in petascale systems

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 a time when the use of GPUs for general purpose computing (known as GPGPU) was still in its infancy. IBM then went on to cancel the PowerXCell project in 2009, and have not returned to accelerator manufacturing.

In the consumer sector, NVIDIA and ATI have been competing for years to dominate the stand-alone GPU market, while Intel controls the built-in solutions. The first petascale system based on GPGPU, Tianhe-1, was launched in 2009 with consumer-grade ATI Radeon graphics cards, while NVIDIA was announcing plans to develop GPGPU-specific devices: the NVIDIA Tesla line of co-processing cards. This was a decisive move for NVIDIA, which has dedicated the most attention to this sector and reaped the rewards: they now control 71% of the accelerated petascale HPC market.

AMD bought ATI in 2006 in a move towards heterogeneous processors with GPU included on the die, which meant no product line would be created to specifically target GPGPU in HPC. Currently AMD/ATI has again appeared in a petascale system with their new FirePro line of GPGPU accelerators used in SANAM, the most energy-efficient of the petascale systems.

The surprising newcomer to the HPC accelerator market is Intel, with their Xeon Phi (previously known as Many Integrated Cores, or MIC) product. This co-processor, used originally in Stampede and then in four more petascale systems including the leading Tianhe-2 computer, is based on a traditional x86 microarchitecture with stream processing. This non-GPU-based co-processor is also very energy efficient, although code must be optimized to use it effectively (as with the GPU-based accelerators).

4.1.4.7. *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), Cray's solutions (SeaStar2+, Gemini, and Aries, now owned by Intel), and IBM's custom interconnects for Blue Gene/P and Blue Gene/Q.

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.

Cray has maintained a high market share (around 20-40%) with their successive generations of interconnects: SeaStar2+, Gemini, and Aries.

The IBM BlueGene IC, which only has two variants (for the Blue Gene/P and BlueGene/Q supercomputer models, has seen a more abrupt rise corresponding with the launch of the two models, and subsequent fall until the next generation.

Fujitsu Tofu and NUDT Arch have a couple of systems each since 2010-2011, and the most important newcomer is Ethernet (in both its Gigabit and 10G versions), which has been growing steadily since 2013 as petascale enters the mainstream (Ethernet has higher latency than the other competitors).

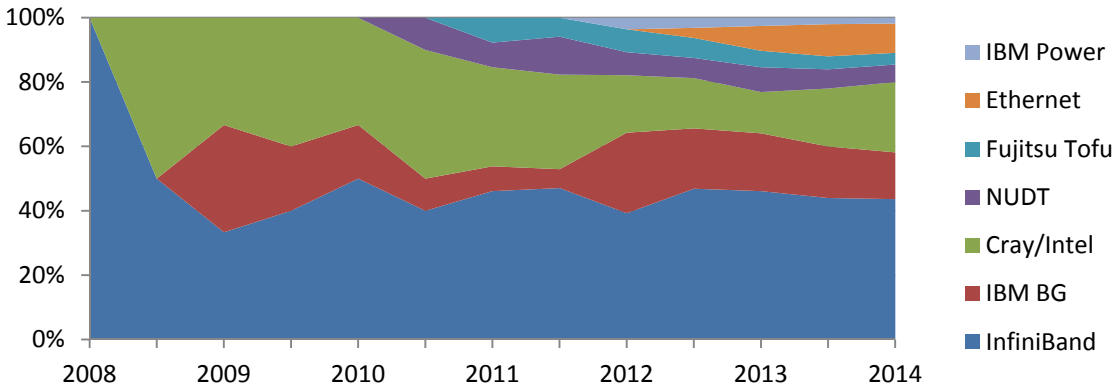


Figure 21: Evolution of interconnects used in petascale systems

4.1.4.8. *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, see Figure 22. 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%.

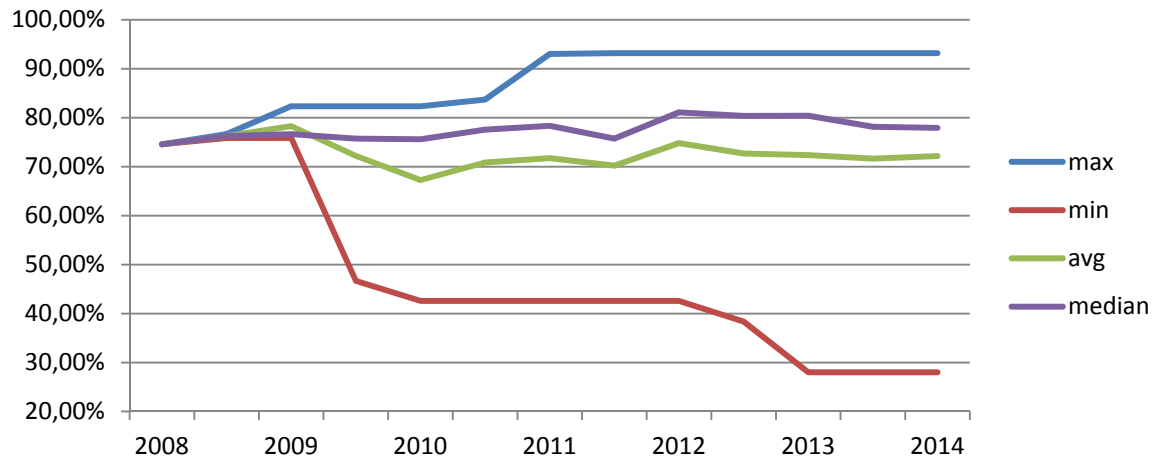


Figure 22: Evolution of the computing efficiency of petascale systems (in %)

4.1.4.9. *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. 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 similar timeframe as that seen earlier based purely on performance. The question is whether the 20 MW limit will stand, or if the desire for exascale will be enough to warrant a higher energy envelope.

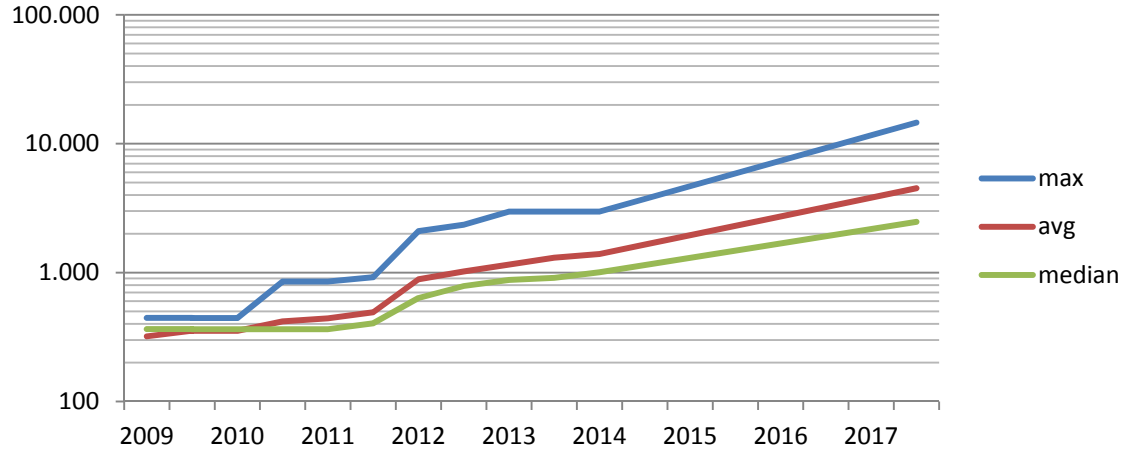


Figure 23: Evolution and prediction (from 2013 onwards) for power efficiency of petascale systems (in MFlop/s/W)

4.1.5 *Beyond HPL*

It is generally accepted that the 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 20 years that allows a 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) reflects just one component of the architecture: the floating-point capability of

the machine, and therefore becomes more and more unreliable as a true measure of system performance for a growing collection of important science and engineering applications (especially those that are reliant on partial differential equations). Today's machines have evolved so that floating-point arithmetic is over-provisioned and runs at incredibly fast speeds, while data movement is becoming the main problem. According to Heroux and Dongarra [4], without some intervention, future architectures targeted toward good HPL performance will not be a good match for real applications.

In 2013 Dongarra, Luszczek, and Heroux presented a new benchmark called HPCG with the goal of striking a balance between floating point and communication bandwidth and latency, to tighten the focus on messaging, memory, and parallelization. At the last ISC'14 event Dongarra presented the first detailed results of the benchmark alongside the Top500 results:

Site	Computer	Cores	HPL Rmax (Pflops)	HPL Rank	HPCG (Pflops)
NSCC / Guangzhou	Tianhe-2 NUDT, Xeon 12C 2.2GHz + Intel Xeon Phi 57C + Custom	3,120,000	33.9	1	.580
RIKEN Advanced Inst for Comp Sci	K computer Fujitsu SPARC64 VIIIfx 8C + Custom	705,024	10.5	4	.427
DOE/OS Oak Ridge Nat Lab	Titan, Cray XK7 AMD 16C + Nvidia Kepler GPU 14C + Custom	560,640	17.6	2	.322
DOE/OS Argonne Nat Lab	Mira BlueGene/Q, Power BQC 16C 1.60GHz + Custom	786,432	8.59	5	.101#
Swiss CSCS	Piz Daint, Cray XC30, Xeon 8C + Nvidia Kepler 14C + Custom	115,984	6.27	6	.099
Leibniz Rechenzentrum	SuperMUC, Intel 8C + IB	147,456	2.90	12	.0833
CEA/TGCC-GENCI	Curie tine nodes Bullx B510 Intel Xeon 8C 2.7 GHz + IB	79,504	1.36	26	.0491
Exploration and Production Eni S.p.A.	HPC2, Intel Xeon 10C 2.8 GHz + Nvidia Kepler 14C + IB	62,640	3.00	11	.0489
DOE/OS L Berkeley Nat Lab	Edison Cray XC30, Intel Xeon 12C 2.4GHz + Custom	132,840	1.65	18	.0439 #
Texas Advanced Computing Center	Stampede, Dell Intel (8c) + Intel Xeon Phi (61c) + IB	78,848	.881*	7	.0161
Meteo France	Beaufix Bullx B710 Intel Xeon 12C 2.7 GHz + IB	24,192	.469 (.467*)	79	.0110
Meteo France	Prolix Bullx B710 Intel Xeon 2.7 GHz 12C + IB	23,760	.464 (.415*)	80	.00998
U of Toulouse	CALMIP Bullx DLC Intel Xeon 10C 2.8 GHz + IB	12,240	.255	184	.00725
Cambridge U	Wilkes, Intel Xeon 6C 2.6 GHz + Nvidia Kepler 14C + IB	3584	.240	201	.00385

These results show that the difference between the systems is in fact less dramatic than shown by the HPL benchmark when local memory system performance and low latency cooperative threading are given significant weight. According to the benchmark authors it reflects customer requirements for vendors much more precisely than LINPACK, and therefore will mean scientific and enterprise benefit, hopefully ensuring acceptance. Even so, they reiterate that it will not replace or diminish the role of the Top500 as an important metric for larger trends in supercomputing (that is, HPCG will always be an addition, not a substitution for HPL).

An additional effort in complementing Top500 list is the Graph 500 list, based on a new set of benchmarks to guide the design of hardware architectures and software systems intended to support data intensive 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). The top 10 systems in the June 2014 edition of Graph 500 can be seen in Table 4:

Rank	System	GTEPS
1	K Computer	17977.1
2	Sequoia	16599
3	Mira	14328
4	JUQUEEN	5848
5	Fermi	2567
6	Tianhe-2	2061.48
7	Turing	1427
8	Blue Joule	1427
9	DiRAC	1427
10	Zumbrota	1427

Table 4: Graph 500 Top 10

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 2014 in Leipzig, Germany).

4.2.1 *Archive storage*

Future exascale systems are expected to generate a lot of data, but even today archiving data is an issue. For example, in many cases researchers require to retain data for a number of years in order to validate their findings. This data does not need to be online and available for calculations, but may be stored on higher latency media if that results in a lower cost. Traditionally this has been tape media, but the increased density of hard drives have made disk based archiving more price competitive.

4.2.1.1. *Disk storage*

Hard drives based on Shingled Magnetic Recording (SMR) are likely to become good candidates for disk based archiving. Since an archive mostly has new data being added and (more rarely) old data being removed, the performance penalty for updating data on a SMR drive is less noticeable.

With 8TB hard drives becoming available, parity with the highest density tape cartridges has been achieved. Having a higher density per unit has traditionally been an advantage for tape.

Erasure coding is commonly used for guarding against data loss due to failing drives and disk enclosures. One example of this exhibited at ISC'14 was the vendor Scality which uses commodity disk servers, in their example HP SL4540, and erasure coding across them build a cluster of disk servers that can sustain multiple failures.

Another potential problem with disks for archiving is having a batch of drives with a manufacturing defect failing at roughly the same time. German vendor FAST LTA is trying to solve this by using disks from multiple vendors mixed in their "storage bricks".

4.2.1.2. *Object Storage*

Usually the term object storage refers to a non-POSIX storage system that does not have the traditional hierarchy of directories and files. Often there is a file system interface provided on top of this, but that is not a requirement. The Lustre file system is an example of object storage with a traditional file system on top. Outside of HPC, the cloud community has been a large user of object storage, there main APIs used are the HTTP REST based S3 (Amazon) and Swift (OpenStack).

Both backups and archiving are really about storing objects, since they are usually not accessed via a file system interface. Archiving was also an early use case for object storage, so in the future we will probably see more backup and archive software using object storage in addition to file systems and tapes. With support for S3/Swift it can also be marketed as "backup to the cloud", which will fit right into current hype trends.

4.2.1.3. *Tape Storage*

A large number of different tape technologies have been created by vendors in the past, and can still be found at sites due to retention requirements and the longevity of tape media. For current larger deployments, the following tape technologies are usually used

- Linear Tape Open (LTO)
- IBM 3592
- Storagetek T10000

LTO systems are available from multiple vendors, while the other two are vendor proprietary.

A critical component in a large installation is the tape library, since a large amount of tape cartridges needs to be mounted and un-mounted without human intervention. LTO based libraries are currently produced by the following vendors

- IBM
- Spectra Logic
- Quantum

Just as in the object storage case mentioned earlier, a file system interface is sometimes added above the tape storage. Traditionally this has been a HSM system, but in recent years most vendors are also providing a LTFS interface.

LTO cartridge capacity has not been keeping up with the increases on other media lately. Between LTO-5 and LTO-6 the uncompressed capacity only increased from 1.5 to 2.5 TB. The current roadmap for LTO has the expected capacity for LTO-7 and LTO-8 at 6.4 and 12.8 TB respectively, but no dates for availability are available. Bandwidth to the tape drive will double between LTO-6 and LTO-7, which will help decrease the time required to utilize a full

tape. To remain price competitive against large SMR hard disks, the price of LTO media will have to be kept down.

When asked about storage density of tape cartridges, some vendors are suggesting their higher density proprietary tape technology as alternative.

4.2.1.4. *File system backup support*

Classical backups require a full scan of the file system being backed up. This is not feasible when dealing with large file systems containing many millions of files and directories. What is needed is some way for the backup client to avoid scanning the file system and comparing it to the list of files currently backed up, but instead obtain the changes since the last backup in some other way.

GPFS from IBM is an example of a file system that provides support for backups via its policy engine. In essence it provides access to the file system metadata allowing the backup application to request a list of changes made since the previous backup was made. IBM provides this integration for their backup software TSM via the mmbackup tool.

Standard Linux file system interfaces like fanotify and inotify does not support network file systems, so currently this requires file system specific integration for backup clients.

4.2.2 *Vendors*

This is a list of vendors which were reviewed by the PRACE team at the ISC'14 conference and exhibition. Although the list is not exhaustive, it gives a good reflection of new trends in HPC.

4.2.2.1. *Bull*

At the beginning of 2014 Bull announced a strategic plan 2014-2017 called One Bull and meant to strongly structure and focus their business around enterprise data and IT security – "trusted partner for enterprise data" was the motto - logically emphasizing Cloud and Big Data sectors. This was followed in May by the announcement of a friendly intended public offer by Atos, a much larger international information technology services company - mostly based in France, Germany, the Netherland, Spain and UK. The merger, fully consistent with the orientation defined by the One Bull plan, is still on-going as of August 2014, and thus no precise comment is possible yet. But this would make the company the Nr 1 European cloud and big data player – more than 80 000 employees all together – with a strong high-end computing technological capacity inherited from the HPC branch.

HPC within Bull has been a steadily growing and flagship activity, although not the largest fraction of their business. A number of "petascale" deals have been made recently by Bull, highlighting dense computing and cooling technological capacities including Direct Liquid Cooling (MeteoFrance, TU Dresden, SURFsara, DKRZ, GENCI-CINES). This has been accompanied by the confirmed development of competency networks and partnerships, such as a reseller partner agreement with Xyratex (beyond the major design win at DKRZ); the further development of Bull's Center for Excellence in Parallel Programming; the further development of Cloud for HPC offer Extreme Factory, incl. participation in the UberCloud initiative.

It will have to be observed, after the end of the acquisition by ATOS, how the HPC "hard core" activity and competency is inserted and valued in the company overall strategy and combined with cloud and big data perspectives.

4.2.2.2. *Cray*

Cray is continuing their supercomputing and big and fast data fusion strategy, with compute products XC30 and CS300; Sonexion, TAS (Tiered Adaptive Storage, connector to Lustre) storage solutions; an incursion into analytics appliances (Urika).

Referring to the Top500 – an arguable business reference but still quite an interesting indicator – Cray is still Nr 3 HPC systems provider with 10% of the total number of systems (50 out of 500, with quite a focus on high-end ones – by contrast with the Top2 providers, resp. HP, focusing more on volume than capability, and IBM, spanning the whole range of system sizes). A pure HPC player, Cray did well in 2014 with quite a number of contracts won worldwide: Korea Meteorological Administration (KMA); Department of Defense High Performance Computing Modernization Program; the Hong Kong Sanatorium and Hospital; the North German Supercomputing Alliance (HLRN), National Energy Research Scientific Computing.

In July it was also announced that Cray will be supplying the DOE/NNSA Trinity machine in two phases with the final phase being complete in 2016. This is one of the largest awards in Cray history – a $174 million deal to provide the National Nuclear Security Administration (NNSA) with a multi-petaflop next generation Cray XC machine, complemented by an 82 petabyte capacity Cray Sonexion storage system.

4.2.2.3. *Fujitsu*

After the K computer, Fujitsu produced the PRIMEHPC FX10 system (2012 – 2015). The Post-FX10 is expected to be “coming soon” (2015). The CPU and interconnect will inherit the K computer architectural concept. The CPU will be a Fujitsu designed SPARC64 XIfx 32 core CPU with some 1 TFlop/s double precision or 2 TFlop/s single precision performance. There will be 1 CPU per node; 12 nodes per 2 U chassis; and some 200 nodes per cabinet. The single CPU/node architecture is a design decision motivated by resulting in a favourable bytes/flop ratio resulting in a better scalability. The system will be water-cooled, and its interconnect will be the next generation Tofu2 which is described in Section 5.1.

4.2.2.4. *HP*

At ISC’14 HP launched the Apollo 8000 and Apollo 6000 systems.

The Apollo 6000 system is designed for best performance per $ per Watt. It is an air-cooled 30 kW/rack solution.

The Apollo 8000 system is designed for leading TFlop/s per rack for accelerated results. The racks are completely warm water-cooled with a special dry-disconnect. A single full rack contains 144 nodes and will be able to deliver 250 TFlop/s using 80 kW. A full rack is 52U high and weights 2132 kg.

NREL, the US National Renewable Energy Lab will use Apollo 8000 for their Peregrine system.

4.2.2.5. *IBM*

One of the most commented HPC announcements of 2014 was obviously the IBM-Lenovo split. About to be fully implemented, the question of whether all IBM customers – like government agencies - will be happy with this situation remains open. This IBM move is however consistent with a vision encompassing and emphasizing global services in the era of

big data and cloud rise, and of efforts of rather developing new learning and data processing technologies such as Watson.

IBM's position with respect to core technologies is evolving as well, as illustrated by the Open POWER initiative. A foundation has been created in 2013 to implement collaboration around Power Architecture products. IBM is opening up processor specifications, firmware and software and is offering this on a liberal license, with the objective to enable the ecosystem players to build their own customized server, networking and storage hardware for future data centres and cloud computing. Google, Mellanox, NVIDIA, Tyan joined the foundation at its early stages. This is however not an ARM-like model yet, and it will be worth observing the evolutions of the model with the next generations of POWER technology design.

Recently IBM has released an announcement pledging 3 billion USD for semiconductor research and core technologies research – future and more miniaturised chips where Moore's Law no longer applies, silicon photonics and carbon nanotubes etc. – see for instance:

http://www.hpcwire.com/2014/07/14/ibm-invests-3-billion-next-gen-chip-design/

This is first a reminder of the strong assets in fundamental research IBM has always fostered and financed in-house - and this will, likely and mostly, not be new money not already spent on research at IBM. But it is a strong assertion of the willingness to invest and bet on mid to long-term research for their product and services development.

4.2.2.6. *NEC*

At SC13 NEC announced its SX-ACE vector supercomputer. End May 2014, the first orders have been announced.

A vector register contains 256 floating-point numbers and has a peak performance of 64 GFlop/s. A node card has one CPU with a total peak performance of 256 GFlop/s and a large memory bandwidth of 64 – 256 GB/s. A single rack consists of 64 nodes (1 CPU per node). A rack has a performance of 16 TFlop/s and a total memory bandwidth of 16 GB/s. A rack uses 30 kW. The interconnect is a dual plane full fat-tree and a node can send/receive 8 GB/s per direction. Available transfer functions are: atomic, 8B. Block sizes are some 32 MB. Furthermore, there are dedicated hardware functions for fast global communications.

4.2.2.7. *RSC Group*

RSC Group is a Russian company founded in 2009. They delivered 40% of the Russian Top500 systems. In 2012 they introduced the RSC Tornado. At SC13 they introduced the RSC PetaStream system. A single (monstrous) direct liquid cooled rack is able to run a world record 1.2 PFlop/s/rack using some 400+ kW/rack. This also boils down to a world record 600 TFlop/s/m^3.

4.2.2.8. *T-Platforms*

T-Platforms is back after a period of being blacklisted by the US government. Measures have been taken to prevent this situation in the future. They installed the June 2014 Top500 #129 at Moscow State University. This is a 420 TFlop/s T-Platforms A-Class system. The high-end A-Class system can be scaled up to 54 PFlop/s. Furthermore T-platforms provides scale out V-Class blade systems and all-in-one T-mini P-Class systems.

4.3 Chapter Summary

The market watch of petascale systems conducted by PRACE since 2010 is providing an interesting view of how the high-end HPC market is evolving. 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.

Accelerators and lightweight cores have both carved a niche in the petascale market, yet neither seem to dominate nor unseat traditional clusters, forcing vendors to spread their offerings to satisfy their client's demands. Clients are also pressuring vendors in relation to application performance, which has led to an open discussion on the validity of High Performance Linpack as a universal measure of computer performance and the introduction of new benchmarks that attempt to solve this increasing disparity.

As petascale enters the mainstream HPC market the number of vendors that offer models with this level of performance is growing substantially, while the "big 2" (IBM and Cray) start to envision what their exascale options will look like. In this regard the recent announcement by IBM of the sale of their x86 server business to Lenovo leaves many doubts about their plans for future top-level HPC systems.

5 Exascalability and Big Data in HPC

Much of the current research into technologies for future HPC systems falls under the exascale umbrella. Likewise, what is currently being called "Big Data" influences the storage systems for traditional HPC clusters, while at the same time starting to require more computational capability. This chapter explores the technology being developed for these fields today.

5.1 CPU, memory and interconnect

5.1.1 *CPU*

A common theme among several less mainstream processor architectures is that they are often developed by engineers with a background in signal processing. This heritage is shown by the tasks targeted initially by these architectures, which often include network processing and video encoding, both suited for digital signal processors.

Due to the requirements for low power consumption for an exascale system, several of the vendors producing these architectures are adapting their products for the HPC market. In one aspect these CPUs are well positioned for HPC usage, as a high-performance interconnect is usually part of a normal implementation, although it is unlikely to be a standard HPC interconnect.

Unless NVIDIA is involved, it can be assumed that OpenCL and OpenMP are the supported programming models for these architectures.

Apart from the examples below, also other many-core processors like the Tilera TILE architecture and the more traditional DSP architectures such as the Texas Instruments Keystone might in the future gain more importance in the HPC market place. While the TI Keystone has fewer cores per socket, it shares many of the same characteristics such as targeting streams of data and being able to be mesh connected. (Previous Keystone coverage in D5.2 [2], section 4.2.6.2.)

5.1.1.1. *OpenPOWER*

Currently the IBM Power architecture is represented by the IBM Blue Gene family (with Q the last) and the IBM pSeries (with IH the top of the line). IBM initiated the OpenPOWER initiative. The OpenPOWER Foundation was announced on August 6, 2013 by Google, IBM, Mellanox, NVIDIA, and Tyan. There are several foundation membership levels. The current platinum members are: Altera, Ubuntu, Google, IBM, Mellanox, Micron, NVIDIA, PowerCore, Samsung and Tyan. Power.org still is the governing body. The basic idea is similar to the ARM situation: IBM / the OpenPOWER foundation will grant licenses. These cover both hardware and software. The current focus is on Power8 but in the future we expect custom Power SoCs. The licensing is not limited to Power8 but might include previous designs as well. For application portability reasons, the OpenPOWER processor architecture will be using little endian mode to match the x86 processor architecture. (Power hardware is well known for being a big endian architecture but is capable of supporting both big and little endian modes.)

Current initiatives include:

- Mellanox exploiting RDMA on Power. Adding native PCIe 3 support without bridges. Mellanox announced a 100 Gbps switch.
- NVIDIA adding CUDA support for NVIDIA GPU accelerators with IBM Power CPUs. NVIDIA will offer its NVLink high-speed GPU interconnect (5 – 12 × PCIe 3) to foundation members.
- Xilinx and Altera FPGA accelerator with CAPI (Coherence Attach Processor Interface) attach to OpenPOWER processors.
- Micron, Samsung, and SK Hynix supplying memory and storage components for an open ecosystem.

5.1.1.2. *Intel Knights Landing*

At the ISC‘14 conference, Raj Hazra from Intel unveiled details of the Knights Landing architecture, including the new Omni Scale fabric.

Knights Landing is the code name of the next generation socket based Intel Xeon Phi (Intel Many Integrated Cores architecture). The Knights Landing processor is expected to be available in the second half of 2015. These processors will be available as PCIe accelerator cards, as well as standalone CPUs. The Knights Landing CPUs will contain Intel Atom cores with a modified version of the Intel Silvermont architecture. Further modifications include AVX-512 and support for 4 threads/core. This results in a threefold thread performance increase versus Intel Knights Corner. A single package will have a 3+ TFlop/s performance. There will also be on-package memory, up to 16 GB at launch. Memory bandwidth and energy efficiency will both be 5 × DDR4.

Intel Omni Scale was announced as the next generation fabric. Its integration will start with Knights Landing and will be incorporated in future 14 nm Intel Xeon processors. Omni Scale will cover a full range of components: integrated, PCIe adapters, edge switches, switch director systems.

On April 29, 2014, NERSC announced the first Knights Landing supercomputer NERSC-8, named Cori. This will be a Cray system and will have more than 9,300 Knights Landing nodes. Its installation is expected to be sometime in mid 2016.

5.1.1.3. *Adapteva Epiphany*

Founded by former Analog Devices DSP engineers and based in the US, Adapteva has some European connection since its current CEO was born in Sweden and the Swedish telecom company Ericsson holds a minority stake of the company.

The current product is the Parallella board that combines 16 or 64 Epiphany cores with two ARM A9 cores. (Technical details were previously covered in more detail in D5.2 [2], section 4.2.6.5.)

At ISC'14 the A-1 reference system was shown. It is a combination of 32 Parallella boards with 64 Epiphany cores each and a 50 Gbit/s proprietary board to board network interconnect. Quoted numbers for power efficiency are 15 GFlop/s/W SP.

A limitation of the current implementation is that it only provides single precision floating point. No products apart from bare development boards are available so far. Thus at the moment this architecture seems to be more at the research stage. Adapteva is collaborating with a number of universities to try and have the architecture used for research in the field of parallel programming.

5.1.1.4. *Kalray MPPA*

Kalray is a French company that was founded in 2008 by a former ST Microelectronics vice president to develop a many-core processor. It started sampling the 256 core Multi-Purpose Processor-Array (MPPA) in 2012, with a second generation processor becoming available this year. In 2014 a new CEO was appointed, and the company now seems to be more focused on the HPC market.

It is a VLIW architecture, and the cores are connected in clusters of 16 cores. One difference from many other competitors is that is does not contain any ARM cores for general-purpose computations. Current implementations include a MMU and will be capable of running a full function operating system. Peak performance is listed as 25 GFlop/s/W, probably SP.

Currently Kalray MPPAs are available as embedded boards used together with x86 host boards. A normal form factor PCIe accelerator card will be available later in 2014. A cartridge for the HP Moonshot is also being developed, but has not been made into a product yet. The PCIe card still requires a host computer, but the Moonshot cartridge will be a standalone computer.

5.1.2 *Memory*

5.1.2.1. *DDR4*

DDR4 memory is the next step in SDRAM memory delivering from 2,133 MT/s to 4,266 MT/s (mega transfers per second) at voltage lowered to 1.2V thus at reduced power consumption. DDR4 will be supported by the Intel Knight's Landing architecture.

5.1.2.2. *NVRAM*

One of the main predicted challenges of the exascale platforms will be the provision of intermediate storage layers in the form of non-volatile random access memory enabling energy efficient caching and transmission of data between processes running on different nodes. Currently several solutions are being developed in this area. The most feasible technologies for this purpose are based on various types of flash storage. However, mostly

due to their possibly cheaper price per gigabyte, several new technologies under development are expected become more suitable in the future:

- *FRAM* – Ferroelectric RAM, memory based on a ferroelectric layer for non-volatile memory storage, which features low power and very high write speeds, although still limited capacity in comparison with other technologies,
- *Racetrack memory* – a novel prototype approach by IBM, has potential to provide much higher data density than existing flash storage solutions with increased performance. The memory is based on the effect of moving magnetic domains in Permalloy nanowire.
- *ReRAM* – Resistive random-access memory is a family of memory technologies using the effect of conduction in the insulating dielectric layer by application of high voltage, which can set or reset a conduction path thus changing the cell value. The main features include less than 10ns switching times, and potentially very high density.
- *PCM* – Phase change memory is a type of NVRAM which exploits the unique behaviour of chalcogenide glass.

5.1.2.3. *Z-RAM*

Zero-capacitor RAM, licensed by AMD, provides parameters comparable to regular DRAM chips; however each bit requires only a single transistor instead of 6, which can greatly increase the memory capacity and overall performance, as well as reduce cost. Z-RAM is expected to be very useful for providing large cache layers between exascale nodes.

5.1.3 *Interconnect*

5.1.3.1. *Mellanox*

At ISC'14 Mellanox unveiled its new EDR InfiniBand switch. It is using silicon photonics. Using 4 lanes, 4 × EDR 25.7 Gb/s it reaches 100 Gb/s transmission speed.

5.1.3.2. *EXTOLL*

The EXTOLL company was founded in 2011 as a spin-off from the University of Heidelberg. The EXTOLL project started in 2005. In 2008 they had the first FPGA prototype and it has been in use since 2009. FPGA-based adapter cards Ventoux and Galibier were introduced in 2011.

The first ASIC implementation of the Tourmalet product was shown at SC13 followed by a PCIe board at ISC'14. The product will be available in Q3 2014.

Tourmalet is designed for HPC:

- >100M messages/s
- Latency < 400 – 600 ns (half round-trip MPI)
- Peak bandwidth >10 GB/s (MPI)
- Core frequency 750 MHz
- HW support multicast, non-coherent shared memory (PGAS), direct device communication (e.g. GPU – GPU)
- Linux kernel drivers, low-level API

EXTOLL is also producing its own connectors and cabling based on Samtec's HDI6 connector:

- Copper: < 1.5m
- Active optical: < 100m
- 120 Gb/s (full duplex)

The topology is a switchless 3D torus direct network limited to 64k nodes. EXTOLL supports fast two-sided messaging and provides an optimized remote direct memory access (RDMA) protocol. Finally, EXTOLL NICs will be fully virtualized.

5.1.3.3. *Fujitsu Tofu2 Interconnect*

The Tofu interconnect was developed for the K computer. The name **To**rus **fu**sion derives from the network topology: a 6D mesh/torus using 4 interfaces with 10 links. The implementation was based on a discrete chip resulting in 40 Gb/s bandwidth.

Tofu2 is designed for Fujitsu's next generation Post-FX10 machine. The implementation will be a SoC and will have 100 Gb/s bandwidth. The most important features will be: atomic Read-Modify-Write operations; cache injection (bypassing main memory) which will reduce latency without additional cache pollution; and offloading of various non-blocking collectives.

5.2 Storage Hardware Development and Basic R&D

The current and especially future HPC systems are useless without efficient storage systems. Storage and distribution of data is a very important element of supercomputers. Computing power is getting bigger and bigger and within this progress the storage system also has to be advanced. Most operations are done in RAM unit, however, the more operations that are done the more input has to be read and the more output data has to be written.

A solution for the increasing requirement for storage performance and capacity are scale-out distributed file systems such a clustered file systems. Even in future projects with innovative ways of data storage as the E10 project, the base of the system will always be a storage medium.

Decision of choosing file system and storage solution for HPC resource is often considered with respect to user's convenience, e.g. file systems shared in all HPC resources installed in a data centre. At present most supercomputers use:

- GPFS
- Lustre – in original or modified version

The decision of choosing storage hardware depends on the budget. Cheap hardware is not reliable and it should be used only with file systems with built-in data protection. File systems protecting from hardware error are:

- GPFS
- Panasas
- SCALI
- StorNext 5 – replication only
- MAHA-FS – replication only

Most of the free of charge file systems lack data protection. Within the last few years only minor progress has been made in this field.

There is always a way to protect data by backup, but it is useless for scratch (work) file systems. Backups can be done only for home directories, which do not have to be as big as scratch and are changed more rarely.

In summary here are three ways of building storage systems for HPC:

- Cheap hardware – complex software
- Expensive hardware – less complex software
- Ready-to-use solutions

This chapter presents information about software and hardware available in the HPC market, including Fraunhofer ITWM, Spectra Logic, Xyratex/Seagate, Panasas, HP, Toshiba, Rausch, IBM, ETRI, BOSTON, Cray, EOFS, Fujitsu, Huawei, Samsung, EchoStreams, GRAU-data, OneStop, Systems, SCALITY, DDN and SanDisk based mainly on information collected during the ISC'14 conference and exhibition.

5.2.1 *Open Source File Systems*

There are two top open source distributed file systems on the market: Lustre and GlusterFS. Lustre is described in section 5.3.1.1.

5.2.1.1. *GlusterFS*

A scale-out and well performing file system that runs on most Linux distributions and comes with features like:

1. Distributed metadata – there is no metadata server – all metadata is stored within a file
2. File based replication
3. InfiniBand RDMA support

Client software uses FUSE and can have performance problems on compute nodes with a high CPU load. This software is getting more and more stable.

5.2.2 *Free of charge*

5.2.2.1. *BeeGFS (former FhGFS)*

The BeeGFS is successfully developed with three key concepts: scalability, flexibility and good usability. The whole software is developed and maintained by a small group of developers without all developers' community overhead.

This file system is simple and performs well without providing as many features as Lustre does. BeeGFS is used in three big HPC installations listed in the June 2014 Top500 list (#22, #56 and #96).

The most noticeable features are:

- Distributed metadata
- Fair I/O algorithms
- Automatic network fail over (IB vs. Ethernet)
- Meta data replication
- Scaling with less than 6% overhead in I/O and throughput compared to a node-local file system

Software is free of charge, but closed source. Paid support is also available.

5.2.3 *Commercial File Systems*

There are many commercial file systems available for HPC use:

- Fujitsu FEFS [5]
- GRAU DATA (Commercial version of OpenArchive, Parallel HSM) [6]

- StorNext 5 [7]
- MAHA-FS [8]
- SCALITY [9]
- GPFS [10]
- Panasas PanFS [11]

5.2.3.1. *Fujitsu FEFS*

FEFS is a Lustre based file system. Compared to the vanilla Lustre version, operations on big storage and big number of files are improved and new features like QoS, directory based quota, jobs I/O zoning and 512KB block size are added.

5.2.3.2. *GRAU DATA*

HSM software with parallel data migration designed especially for Big Data. Runs with Lustre v2.5.

The open source version of GRAU DATA is developed without parallel migration as OpenArchive software.

5.2.3.3. *StorNext 5*

Created as a distributed file system with FC interconnect to clients. Now clients support also Ethernet interconnect. StorNext 5 is a full HSM system.

5.2.3.4. *MAHA-FS*

MAHA-FS is a file system created by Korean ETRI. The software is based on GLORY-FS. It is said to be more reliable than Lustre with data tiering support but not as good performance and scaling behaviour as Lustre. It is a compromise between hard drive cost and performance.

5.2.3.5. *SCALITY*

The SCALITY is introduced as software defined storage. This software runs on all common servers. The advantages of this software are:

- Object Storage technology
- Reliability, high performance and low latency
- Advanced Resilience Configuration (ARC) data protection
- Multi-site data protection
- Multi protocol file access

The SCALITY software is a promising technology, with licensing paid per TB independent of the number of connected clients. The Linux client software uses FUSE, which can cause performance problems on CPU-intensive computing nodes. Performance tests are needed before using this product in a HPC cluster.

5.2.3.6. *GPFS*

GPFS is a well-known and widely used file system from IBM, which is well established in the HPC market. Features include GPFS Native RAID (GNR) technology dedicated for fault tolerant large-scale storage systems with up to 100,000 disk drives.

5.2.4 *Future projects*

Exascale 10 (E10) is a project part-financed in the DEEP-ER and the Mon-Blanc2 EU projects. The main objective is to develop a ubiquitous middleware that helps I/O operations in the Big Data regime aside from file system bottlenecks; data should be placed directly on storage by intelligent middleware, without serialisation and with respect to hardware configuration such as stripe size.

E10 assumes non-POSIX storage access, which means that access to storage has to be implemented in the applications. The benefit of using the E10 solution is good performance and data storage setup facility.

5.2.5 *New hardware mediums for Big Data*

5.2.5.1. *Hybrid HDD*

Hybrid drives were introduced on the market more than a year ago. They are traditional HDDs with NAND memory added as accelerator for data storage. Thus, using sophisticated algorithms for block migration, it gives performance close to SSD drives.

Problematic is endurance of NAND memory, which is write destructive. If most input operations go through the accelerator the lifetime of such drive can be short.

Most of these types of drives are not recommended for HPC, where most I/O operations are writes. Some vendors sell hybrid drives only for personal computers.

5.2.5.2. *NAND Storage in DIMMs or PCIe*

Installation of NAND memory in DIMM or PCIe slots is getting more and more popular.

An example specification for such a solutions is:

- 2-5 µ s latency
- 1GB/s read 750GB/s write
- 60-150KIOPS / DIMM
- 200/400 GB / DIMM

Such specification makes this solution unbeatable in performance. When using PCIe Expander 8mln IOPS, 40GB/s with 100TB capacity can be achieved from 3U chassis (One Stop Systems Flash Array product).

5.2.5.3. *V-NAND*

Vertical NAND are devices with innovative architecture of memory cells which are stacked vertically. The vertical layers allow larger bit concentration without cells downsizing. Table 5 shows the options for connecting it to the host system.

Samsung is the only vendor offering memory built with this technology.

Interconnect	**SATA III**	**PCIe**	**SAS**
Capacity [GB]	960	800 / 1600 / 3200	400 / 800 / 1600
Read/write [MB/s]	530 / 430	3000 / 1400	1400 / 1100
Read/write [kIOPS]	89 / 35	750 / 200	200 / 50
DWPD in 5 years	3.5	10	10
Power [W]	2.6	25	9

Table 5: V-NAND connectivity

5.2.5.4. *HDD with Ethernet connector*

An innovative way to connect HDDs directly via Ethernet to other IT systems. Each disk drive has a built-in Ethernet controller with:

- DHCP support
- 1GE Interface (10GE – Q4 2014)

There is a need to install software for consolidation of such drives like *BigFoot Storage Object.* This hardware is not dedicated for HPC use; its use is mainly recommended for web-services.

5.2.6 *Storage Systems Solutions (Hardware + Software)*

There are ready-to-use storage systems with preinstalled software that exports file systems. There are two groups of such systems: with Lustre software or with proprietary file system software.

5.2.6.1. *Based on Lustre*

All hardware configurations support the following features: Failover configuration for metadata server (MDS) and object storage targets (OSTs), hardware optimizations for the Lustre software, and an easy to use and convenient interface with full hardware monitoring. Major vendors are:

- DataDirect Networks: more than 3.5GB/s from a single client, performance scales to TB/s
- EMC2: Lustre with flash tier (aBBa – Active Burst Buffer Appliance), HSM appliance for Lustre and Isilon with GRAU Software
- Xyratex (ClusterStor): 1.6PB / rack, up to 42GB/s per rack
- Cray: Xyratex Cluster Stor OEM with CRAY support

5.2.6.2. *Proprietary*

Panasas

Appliance with the PanFS file system with features like:

- Data access with protocol PanFS, pNFS, NFS and CIFS
- License per client per TB
- Support Hybrid Storage Blades – tiering support with SSD drives.
- Triple parity data protection – with faster than traditional RAID 6 rebuild
- Per-file data protection
- Scale over 12PB, 150 GB/s and 1.3M IOPS
- Automated load-balancing
- InfiniBand support

StorNext

This is an appliance using the StorNext 5 file system. Clients are connected via FC or GE or IP over InfiniBand. Storage appliance is equipped with FC ports only. For data access with non-FC medium a special type of appliance is needed.

5.2.6.3. *Summary*

The most promising medium for the future is the NAND based memory in the SSD or PCIe or DIMMs format. This type of memory is getting more and more reliable with better endurance during the write process.

There are plenty of scale-out file systems ready to use, from mature Lustre and GPFS with many developers and a huge installation base, to simple and fast BeeGFS with just a few developers and few installations worldwide.

The Lustre file system is in a good condition, with both Intel and Xyratex (now Seagate) being involved in the software development. Most Storage Systems solutions for HPC are based on Lustre.

New promising file systems are appearing, e.g.: SCALE, BeeGFS, and IBM GPFS on Hadoop.

Systems such as Lustre, SCALITY, GPFS, MAHA support file tiering (file migration between different disk groups).

5.3 Storage Organization for Exascale Computing Systems

The data management problem in HPC involves generally two aspects. One is the IO data transfer from the computing nodes to the local network attached storage. The second one involves data transfers between different computing machines, possibly in different institutes distributed geographically. As for the first challenge some mitigation is possible, for instance through in-situ processing approaches, the second case will see even larger disproportion in exascale and will require substantial advanced planning and prefetching in order to ensure that required data is available near the computing nodes where the computation is taking place, or scheduling the computation to machines where the data already is available.

The main challenges related to data management systems in exascale machines will include [12]:

- **Concurrency** – substantial increase in the number of cores performing parallel processing and performing independent I/O operations will make current I/O approaches impossible due to enormous disproportion between the increase in the number of cores per machine and the increase in network throughput,
- **Data locality** – while the number of cores per processor/node increases significantly it will be crucial for the computing system to provide a temporary storage accessible directly via bus which can be used for storing temporary results and improve in-situ processing. This obviously raises several issues with data synchronization between different threads/processes and requires significant revision of existing algorithms in order to minimize synchronization between different threads executing in nodes connected through "slow" network interconnects,
- **Slow I/O** – the current trend shows clearly that even increasingly more efficient network interconnects and storage access times will not be even close to allow exascale machines to use them as "local" storage which could be used for temporary results. This challenge can be addressed through modification of algorithms towards in-situ approaches; however this cannot always be done. The problem remains of what can be done for applications that will require a constant stream of I/O operations on network attached storage devices. This problem can only be minimized through provision of a truly high performance parallel file system, with multiple independent interconnects between the computing nodes and storage allowing advanced algorithms

to schedule task execution in certain topologies minimizing any bottlenecks on single connections.

The current architectures which split the HPC system into compute nodes and network attached storage will not be suitable for exascale machines, where the number of cores and their peak throughput will significantly outscale the network throughput to the local network storage. One solution is to provide each core/processor/node with a high performance medium capacity local storage supporting the in-situ algorithms and thus minimizing the network bandwidth to the common storage. Thus the main work within this project will include addressing the data management and I/O issues in exascale systems on several levels.

Many problems of science, engineering, etc. are getting increasingly complex and interdisciplinary due to multiple data sources and computational methods. A common feature across many of these problem domains is the amount and diversity of data [13]. Such Big Data is increasingly large-scale and distributed, arising from sensors, scientific instruments, simulations, and storage clouds. Applications need access to these data-sets and to the facilities that handle them. However, as data processing becomes a larger part of the whole problem, either in terms of data size or data-mining/processing/analytics, new paradigms are becoming important. For example, most data analytics involves linear algebra or graph algorithms. Furthermore, storage and access to the data naturally involves database and distributed file systems as an integral part of the problem. It has also been found that much data processing is less closely coupled than traditional simulations and is often suitable for distributed dataflow based software runtime systems.

It is expected [14] that exascale systems will have about 10^2-10^3 fold increase in system memory while 10^5 increase in concurrency. This means that typical applications will be divided into many threads, requiring substantially larger I/O throughput in order to allow relevant scaling and performance in hundreds of PFlop/s. Existing data management technologies and high performance file systems do not have this kind of capacity, and are not expected to improve in a rate sufficient to sustain throughput requirements of exascale massively parallel applications.

In a Department of Energy commissioned report [15] the authors have compared the existing petascale systems metrics with estimates for future exascale machines. Data from this study is shown in Table 6. With respect to storage the increase is 20-fold in terms of storage space, 30-fold increase in interconnecting bandwidth and almost 100-fold in terms of I/O bandwidth necessary to sustain exascale computations. Existing solutions cannot be scaled to such a degree.

Feature	2010	2018	Factor change
System peak	2 Pflop/s	1 Eflop/s	500
Power	6 MW	20 MW	3
System memory	0.3 PB	10 PB	33
Node Performance	0.125 Gflop/s	10 Tflop/s	80
Node Memory BW	25 GB/s	400 GB/s	16
Node Concurrency	12 cpus	1000 cpus	83
Interconnect BW	1.5 GB/s	50 GB/s	33
System Size (nodes)	20K nodes	1M nodes	50
Total Concurrency	225K	1B	4444
Storage	**15 PB**	**300 PB**	**20**
IO bandwidth	**0.2 TB/s**	**20 TB/s**	**100**

Table 6: Comparison of petascale vs. exascale system performance

Currently, we can identify 4 basic types of storage architecture with respect to HPC computing nodes:

- Computing nodes with a distributed file system (e.g. Lustre) and separate data management solutions between sites (e.g. DPM)
- Computing nodes with a global data management system (e.g. iRODS)
- Computing nodes with local storage (e.g. SSD) and a global data management system
- Computing nodes with local storage along with a distributed file system and data management system

Among these, with respect to exascale systems, most important seem to be the ones that include a substantial amount of local storage, i.e. storage available directly at the computing node. It is important however, to distinguish this storage from a simple local node disk with an operating system, as this local disk has to be part of the file system namespace and data generated by tasks running on these computing nodes must be available to other computing nodes on demand during task execution, without requiring network transfer outside of the compute cluster. This approach ensures that in-situ applications can gain maximum performance improvement from their architecture.

The following subsections mention existing potentially exascale scalable solutions.

5.3.1.1. *Lustre*

In 2012 Lustre [16] won the FastForward Department of Energy contract for enhancing Lustre implementation to support exascale systems in the future under the name of Exascale Fast Forward IO stack (EFF) [17]. Its development for exascale includes such aspects as extended support for HDF5 at the application level, I/O dispatch layer (called Burst Buffer) optimizing transfer between application and storage layers and Distributed Application Object Storage (DAOS) supporting multi-version concurrency control. Due to analysis of cost vs. performance and capacity ratios, it is proposed that the middle layer can be based solely on SSD storage while the bottom DAOS layer should be a hybrid solution.

Figure 24 presents an overview of the EFF deployment where user applications running on compute nodes access the system mainly through the top-level Mercury I/O forwarding API, which then passes the IO operations to the IO Dispatcher layer which performs transfer optimization between top and bottom layers.

Lustre is also the basis of an Exascaler software/hardware solution [18] developed by DataDirect Networks. The solution scales to tens of PBs of storage and achieves throughput per client at the level of 3.5GB/s on InfiniBand.

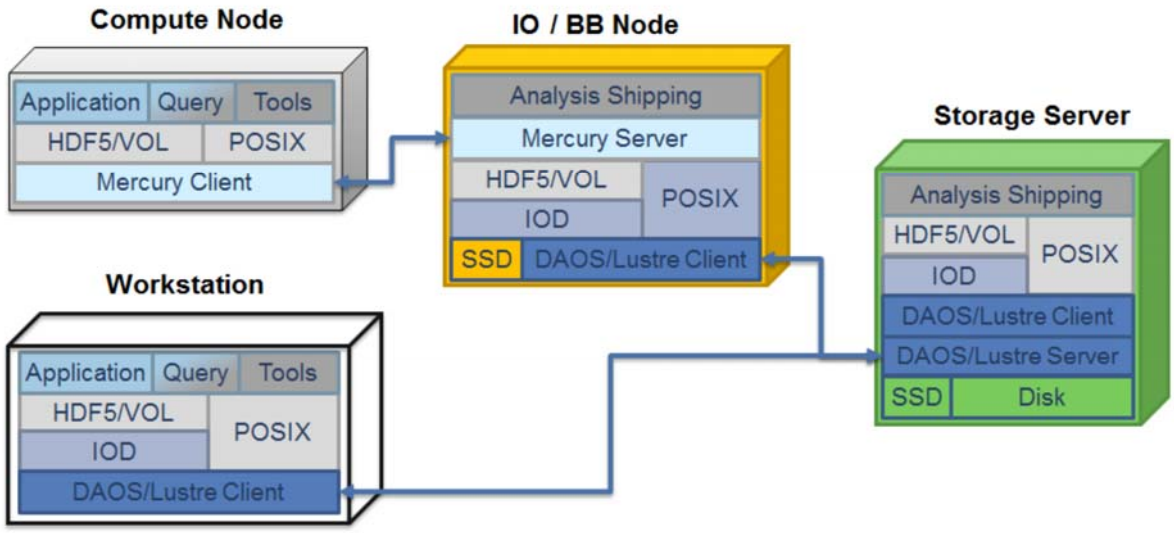


Figure 24 Exascale Fast Forward (EFF) I/O stack prototype [17]

5.3.1.2. *GPFS/Elastic Storage*

The General Parallel File System (GPFS) is an IBM product [10], recently rebranded to Elastic Storage. The main performance feature of GPFS is the striping of files into blocks (which can be even smaller than 1MB) and optimizing data placement on such basis, thus enabling significant performance improvement when adding more storage capacity. Furthermore, GPFS provides support for metadata distribution, distributed locking enabling POSIX compliance and recovery from partition failures.

Although GPFS already supports very high-bandwidth operation (for instance the Jülich installation manages 7PB of data at 200GB/s bandwidth [19]), it is not clear how its locking mechanism would scale to exascale platforms [20].

Within the framework of the Exascale Innovation Center (a collaboration between Jülich Supercomputing Centre and IBM), GPFS was evaluated for the possibility of managing NVRAM storage cache [21].

5.3.1.3. *EIOW (Exascale IO Workgroup)/EOFS (European Open File System)*

EIOW [22] is a workgroup within the framework of the European Open File System non-profit organization, which aims at providing architecture and guidelines for enabling exascale storage in the future HPC platforms. In their architecture whitepaper [23], the workgroup proposes to take a 2-step approach to the exascale storage solutions by first providing additional tools and interfaces leveraging existing parallel file systems such as Lustre or GPFS (see Figure 25), and building in parallel core storage solution optimized for exascale.

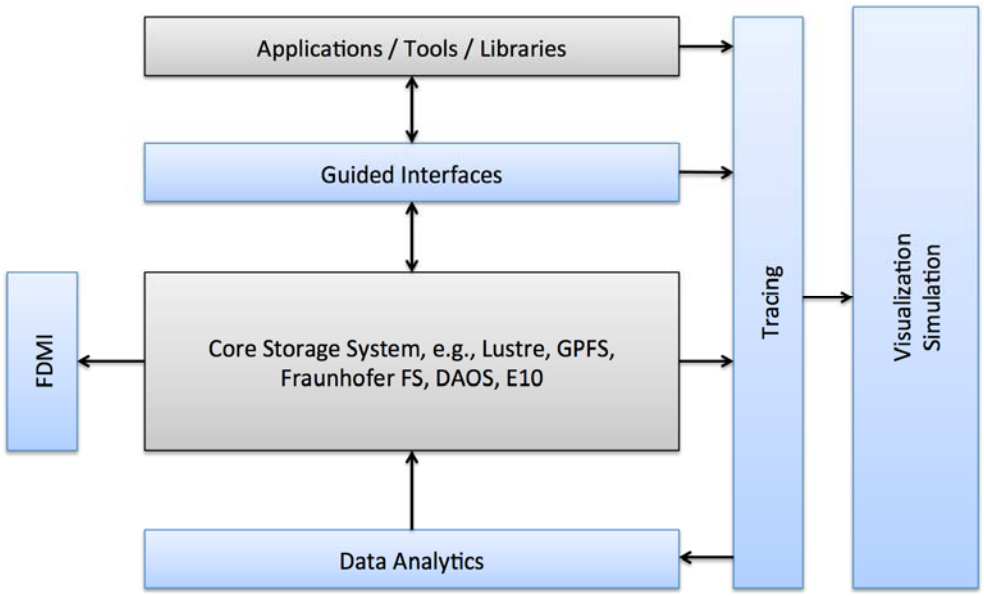


Figure 25 E10 exascale storage architecture

5.3.1.4. *Infinite Memory Engine*

IME [24] is a product by DataDirect Networks which provides a so-called burst-buffer cache for significantly improving the performance of existing underlying parallel file systems such as Lustre or GPFS, and enabling potential scalability to exascale platforms. The main feature promised by this solution is significantly reduced check-pointing time, enabled by the distributed NVRAM cache layer and custom algorithms minimizing the locking overhead.

5.3.1.5. *Other*

Some proposed solutions to the exascale data management issues have been presented already. [25] proposes a solution to the storage access latency problem, called DASH. It introduces Flash based I/O super nodes connected with InfiniBand.

In [20], authors give a significant motivation for the need of major redesign of existing approaches to storage management and file systems. They propose a novel file system solution, called FusionFS, which provides a standard user level POSIX interface while handling novel storage architecture where computing nodes have local storage which can be shared between other nodes.

5.4 EU Projects for Exascale and Big Data

The timescale for demonstrating the world's first exascale system is estimated to be 2018-2020 (a variety of estimates exist). 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/MIC designs)
- High-bandwidth, low-latency novel topology networks (almost certainly custom-designed)
- Faster, larger, lower-powered memory modules (perhaps with evolved memory access interfaces).

Deliverable D5.2 [2] described three exascale related EU projects: DEEP, Mont-Blanc and CRESTA. A few additional projects started in 4Q2013: DEEP-ER, EPiGRAM, EXA2CT, NUMEXAS and Mont-Blanc 2. The H4H described in this section was a project funded by national ministries of industry and research in France, Spain and Germany. H4H is not an exascale project, but was delivering several advanced tools and application framework for future technologies and HPC architectures.

All of them are trying to propose improvements for exascale systems on various levels:

- Hardware improvements – proposition of new architectures
- Middleware improvements – new tools, programing libraries
- Working out new algorithms for exascale systems – software libraries, new simulation techniques.

The description of these projects in the following subsections is largely extracted from the respective project web sites.

5.4.1.1. *DEEP*

The DEEP – **Dynamical Exascale Entry Platform – project** (http://www.deep-project.eu/) is an exascale project funded by the European 7th FP. The DEEP project 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 final DEEP prototype system will consist of a 128 node Eurotech Aurora Cluster and a 512 node Booster.

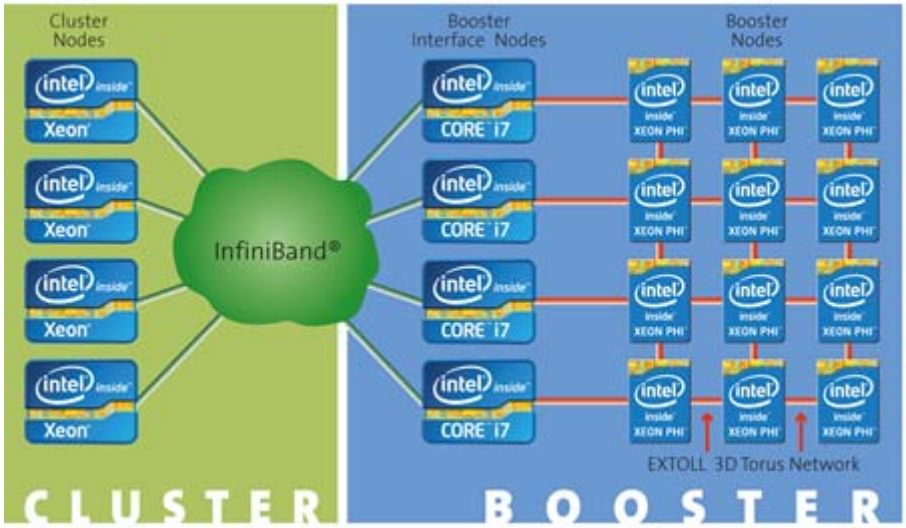


Figure 26 DEEP hardware architecture

The DEEP programming model provides a dedicated development and runtime environment (Figure 27):

- If the application supports MPI, it will benefit from the optimized MPI implementations for the InfiniBand fabric on the Cluster side as well as for the EXTOLL network on the Booster side.

- If the application does not use MPI, the task-based OmpSs programming model will support the application run

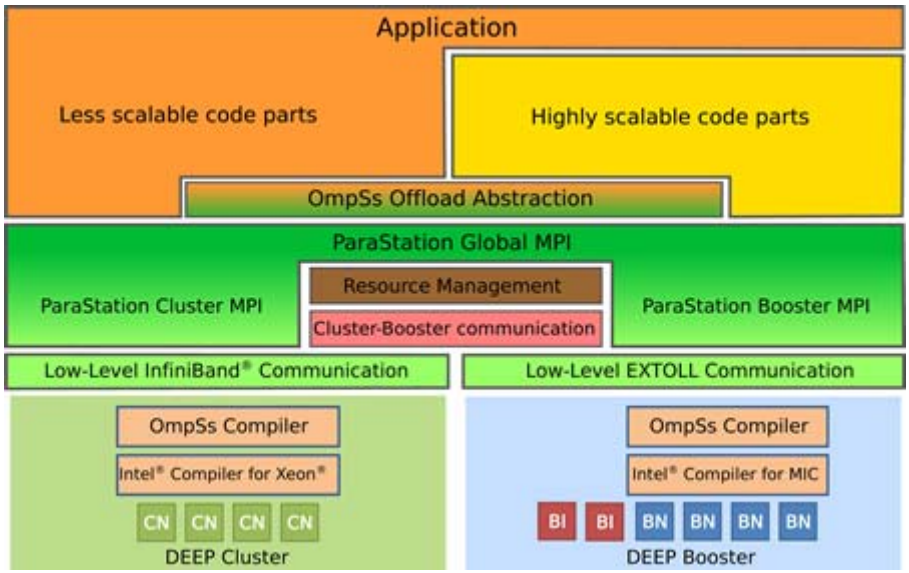


Figure 27: DEEP software architecture

5.4.1.2. *DEEP-ER*

The ***Dynamic Exascale Entry Platform – extended Reach*** project (DEEP-ER, http://www.deep-er.eu/) aims at extending the Cluster-Booster Architecture that has been developed within the DEEP project with a highly scalable, efficient, easy-to-use parallel I/O system and resiliency mechanisms. A Prototype will be constructed focusing on hardware components and integrating new storage technologies. They will be the basis to develop a highly scalable, efficient and user-friendly parallel I/O system tailored to HPC applications. Building on this I/O functionality a unified user-level checkpointing system with reduced overhead will be developed, exploiting multiple levels of storage.

The DEEP programming model will be extended to introduce easy-to-use annotations to control checkpointing and to combine automatic re-execution of failed tasks and recovery of long-running tasks from multi-level checkpoint. The requirements of HPC codes with regards to I/O and resiliency will guide the design of the DEEP-ER hardware and software components. Seven applications will be optimised for the DEEP-ER Prototype to demonstrate and validate the benefits of the DEEP-ER extensions to the Cluster-Booster Architecture.

The DEEP-ER project will feature a highly scalable, efficient, and user-friendly parallel I/O system based on the Fraunhofer parallel file system BeeGFS (formerly known as FhGFS). Additionally, it will provide a low overhead, unified user-level checkpointing system and exploiting the multiple levels of non-volatile memory and storage added to the DEEP architecture.

The DEEP-ER project will focus on two features:

- Resiliency

  Exascale systems will require powerful resiliency techniques that are also flexible enough to accommodate the heterogeneous nature of systems like the DEEP and DEEP-ER prototypes. Isolating partial system failures to avoid full application restarts will be a key to allow compute at the exascale

- Scalable and parallel I/O system

  Highly scalable and parallel I/O will be an important building block for future exascale systems. In the DEEP-ER project, the I/O subsystem relies on 3 components: Fraunhofer's parallel file system BeeGFS, the parallel I/O library SIONlib, and Exascale10.

The design process features system components that can be upgraded with better implementations or even newer technology. Apart from that, the start-up costs and risks involved in producing highly specialized boards are avoided. This allows the use of second-generation Intel® Xeon Phi™ many-core CPUs that boot without the help of an attached Intel® Xeon® processor for the Booster part. Due to this, DEEP-ER will be able to use the same interconnect network spanning both Cluster and Booster, while the DEEP prototype is based on two distinct networks.

Additionally, to support highly efficient I/O and fast checkpoint/restart systems, DEEP-ER attaches novel, non-volatile memory to the Booster nodes, and it evaluates the concept of Network Attached Memory (NAM) as a shared, persistent memory resource.

5.4.1.3. *Mont-Blanc/Mont-Blanc 2*

The Mont-Blanc project (EUROPEAN APPROACH TOWARDS ENERGY EFFICIENT HIGH PERFORMANCE, http://www.montblanc-project.eu/) started in October 2011 and will last until September 2014.

A new project called **Mont-Blanc 2, European scalable and power efficient HPC platform based on low-power embedded technology** (October 2011 – September 2016) has already started and is aiming to:

- Support Mont-Blanc
- Complement Mont-Blanc with system software stack
- Evaluate ARM-based platforms
- Define future Mont-Blanc exascale architecture.

Energy efficiency is already a primary concern for the design of any computer system, and it is unanimously recognized that future exascale systems will be strongly constrained by their power consumption. This is why the Mont-Blanc project 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 times less energy.

Mont-Blanc 2 contributes to the development of extreme scale energy-efficient platforms, with potential for exascale computing, addressing the challenges of massive parallelism, heterogeneous computing, and resiliency.

5.4.1.4. *CRESTA*

The CRESTA project (Collaborative Research into Exascale Systemware, Tools & Applications, http://cresta-project.eu/) is a FP7 EU project with the aim to provide exascale requirements from the end user point of view and new tools and systemware.

CRESTA brings together four of Europe's leading supercomputing centres, with one of the world's major equipment vendors, two of Europe's leading programming tools providers and six application and problem owners to explore how the exaflop challenge can be met.

The project has two integrated strands: one focused on enabling a key set of co-design applications for exascale, the other focused on building and exploring appropriate systemware for exascale platforms.

CRESTA's key objectives are:

- To build and explore appropriate systemware for exascale platforms. From the development environment, through algorithms and libraries, user tools, and the underpinning and cross-cutting technologies, it aims to produce an integrated suite of systemware to progress European competitiveness.
- To enable a set of key co-design applications for exascale. Representing an exceptional group of applications used by European academia and industry to solve critical grand challenge issues, it will enable these to prepare for and exploit exascale technology, enabling Europe to be at the forefront of solving world-class science challenges.
- Co-design is at the heart of the project. CRESTA's objective is to ensure the co-design applications provide guidance and feedback to the systemware development process and integrate and benefit from this development in a cyclical manner.
- CRESTA aims to use a dual pathway to exascale solutions: employing incremental and disruptive approaches to technical innovation – sometimes following both paths for a particular problem to compare and contrast the challenges associated with each approach.

The following recommendations were made in the newly released White Paper[4] (The Exascale Development Environment – State of the art and gap analysis):

- **Tools Integration in Scalable Framework**: Tool integration should be developed. Allinea's tools (http://www.allinea.com/) represent a scalable portable platform that will be enhanced to reach exascale and this platform can be opened as a way to provide a lower barrier to entry for other tools developers, removing the hard problems of scalability and portability and allowing them to concentrate on their strengths.
- **Support for New Programming Models**: As other elements of CRESTA will investigate programming models, the impact of debugability is important, and an alternative model will be identified and debugging support for this will be considered with a view to discovering how such models will be debugged.
- **MPI Correctness**: While an interesting paradigm/model for an extension of existing correctness checkers should be identified, MPI remains a primary interest for application developers. As a result, extending the scalability of MUST to cope with more than 1,000 processes is the projects priority. This involves extensions for runtime message matching and deadlock detection.
- **Clustered Anomaly Detection**: Debugging is both deductive and iterative. At current scale, and as it reaches higher scales, it can automatically identify anomalies that happen; differences to previous successful runs, and with processes that are successful within the current task. This could cover both data changes, and process activity. Automated methods for asserting data integrity should be investigated that would allow, for example, a developer to efficiently detect incorrect values. This could involve developing both standard libraries for data verification, and model specific libraries.
- **Application/Library Model Awareness**: Better integration of layered models and the debugger should be investigated. For example, awareness of MPI communicators and the internals of request object or integration with runtime of task based parallel frameworks to visualise internal task lists.

[4] http://cresta-project.eu/images/cresta_whitepaper.pdf

5.4.1.5. *EPiGRAM*

The Exascale ProGRAmming Models (EPiGRAM, http://www.epigram-project.eu/) project is an EC-funded FP7 project on exascale computing. The aim of the EPiGRAM project is to prepare Message Passing and PGAS programming models for exascale systems by fundamentally addressing their main current limitations. The concepts developed will be tested and guided by two applications in the engineering and space weather domains chosen from the suite of codes in current EC exascale projects.

Major project objectives are:

1) EXASCALE MPI: The project will investigate innovative and disruptive concepts in the MP programming model, and implement them to tackle the challenge of MP scalability on exascale computer systems. Especially important is to investigate how MPI (or better: an MPI-like message-passing model/interface) can coexist with other models, like PGAS. This will be highly useful for application programmers working at extreme scale, and will be implementable with high-efficiency on exascale system.
2) EXASCALE PGAS: The project will first investigate disruptive concepts in PGAS programming models, such as scalable collective operations based on the one-sided communication model, and improved synchronization mechanisms, and then implement them in GPI. It will investigate fault tolerance strategies in the PGAS programming model and then implement them in GPI. It will try out different methods for the interaction between the communication library and the application, like notifications of the application about suspicious resources or timeouts when trying to communicate to faulty nodes. To allow libraries to be executed in a separated communication domain, it will investigate implementations of segmentation of memory and dynamic allocation of resources like communication queues and implement them in GPI.
3) Programming models for diverse memory spaces: The project will first investigate the state of the art in memory-hierarchy aware communications models and implementations. This information will be communicated to relevant standards bodies, with suggestions and proposals where appropriate. It will explore the most efficient ways to use available communications models and libraries (especially those developed in EPiGRAM) in HPC systems with hierarchical memory models. This will be done using appropriate benchmark codes, representative EPiGRAM application kernels and the full EPiGRAM applications.
4) Exascale PGAS-based MPI: The project will bridge the gap between the two approaches by investigating the two programming models where the appropriate constructs are used depending on the requirements of the application. This project will determine the necessary pre-requisites needed to support these hybrid programming models at extreme scale and demonstrate the viability of this approach by producing open source software libraries to efficiently support these hybrid models.
5) Exascale-ready applications: The project will use the exascale MP implementation and PGAS libraries in two real-world applications, Nek5000 and iPIC3D, to prepare them for exascale. It will develop new exascale communication kernels in Nek5000 and iPIC3D with the goal of achieving high scalability and enabling new science to be carried out.

The testbed is using two applications: **Nek5000** (FORTRAN and MPI simulation code for the simulation of incompressible flows in complex geometries. The code is used to study the fluid

dynamics of fission nuclear reactors and improve their design in order to avoid accidents) and **iPIC3D i**s a C++ and MPI Particle-in-Cell code for the simulation of space and fusion plasmas. The code is used to simulate the interaction of solar wind and solar storms with Earth magnetosphere and spacecrafts, and the plasma in magnetic confinement fusion devices, such as Tokamaks and reversed field pinch machines.

The EPiGRAM project started November 2013. It is coordinated by KTH (Sweden).

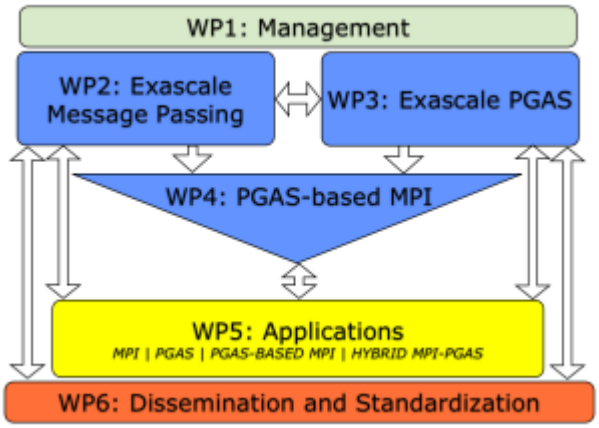


Figure 28 Dependencies between work packages

5.4.1.6. *EXA2CT*

The ***EXascale Algorithms and Advanced Computational Techniques*** project (EXA2CT, https://projects.imec.be/exa2ct/)

The EXA2CT project brings together experts at the cutting edge of the development of solvers, related algorithmic techniques, and HPC software architects for programming models and communication. It will take a revolutionary approach to exascale solvers and programming models, rather than an incremental approach. It will produce modular open source proto-applications that demonstrate the algorithms and programming techniques developed in the project, to help boot-strap the creation of genuine exascale codes.

The goal of this project is to develop novel algorithms and programming models to tackle what will otherwise be a series of major obstacles to using a crucial component of many scientific codes at exascale, namely solvers and their constituents. The results of this work will be combined in running programs that demonstrate the application-targeted use of these algorithms and programming models in the form of proto-applications. The application targeting will be done by an analysis of a representative selection of scientific applications using solvers and/or the constituent parts that the project targets. The results of the project will be disseminated to the reference application owners through a scientific and industrial board (SIB), and board-partner specific code targeting activities, to help generate momentum behind our approach in the HPC community. The proto-applications will serve as a proof-of-concept, a benchmark for doing machine/software co-design, and as a basis for constructing future exascale full applications. In addition, the use of the SIB is a means to extract the commonalities of a range of HPC problems from different scientific domains and different industrial sectors to be able to concentrate on maximising the impact of the project by improving precisely those parts that are common across different simulation needs.

The main objectives of EXA2CT are:

- Discover solver algorithms that can scale to the huge numbers of nodes at exascale.

- Develop highly scalable preconditioning relevant to many industrial and scientific codes.
- Develop an exascale programming model that is usable by application developers.
- Support the algorithmic work, improve the applicability of highly efficient stencil compilers to a wider class of problems. This will be verified by constructing kernel examples from the expanded problem class.
- Enable efficient codes for exascale machines, both at the node and cluster levels, by bringing together a library with distributed data structures, stencil compilers and an efficient on-node task scheduling system.
- Research and demonstrate algorithm level resilience for linear solvers to help alleviate the constraints related to check-pointing, and to link this to error detection in the runtime environment and recovery procedures.
- Offer these developments to the wider community in open-source proto-applications, to enable exascale machine/software co-design and a basis for exascale applications.
- Target the proto-applications to the needs of the board members by working with them.

5.4.1.7. *NUMEXAS*

The NUMEXAS project (***Numerical Methods and Tools for Key Exascale Computing Challenges in Engineering and Applied Sciences***, http://www.numexas.eu/) started in October 2013 and will last for 36 months.

Numexas, "Numerical Methods and Tools for Key Exascale Computing Challenges in Engineering and Applied Sciences", is a STREP collaborative project within the FP7-ICT programme of the European Union. The goal of Numexas is to develop, implement and validate the next generation of numerical methods running on exascale computing architectures. It will achieve this goal by implementing a new paradigm for the development of advanced numerical methods that is able to fully exploit the intrinsic capabilities of the future exascale computing infrastructures.

The main outcome of Numexas is a new set of numerical methods and codes that will allow industry, government and academia to solve exascale-class problems in engineering and applied sciences on the next generation of exaflop computers, with the efficiency and ease of use as today's state-of-the-art codes.

The Numexas consortium includes renowned institutions specialised in the development of numerical methods to solve scientific and engineering problems: CIMNE, the International Center for Numerical Methods in Engineering and coordinator of the project, the group IKM of the LUH, the Gottfried Wilhelm Leibniz Universitaet Hannover in Germany and the National Technical University of Athens in Greece and institutions hosting HPC facilities and supercomputing infrastructures (the Consorci Centre de Serveis Cientifics i Academics de Catalunya, CESCA in Spain and the group HKNR in LUH). The partnership is completed with QUANTECH, an SME specialised in the development and marketing of simulation software for industrial forming processes.

The overall aim of NUMEXAS is therefore to develop, implement and demonstrate the next generation of numerical simulation methods to be run under exascale computing architectures. This cannot be done by just scaling currently available codes, but by implementing a new paradigm for the development of advanced numerical methods to really exploit the intrinsic capabilities of the future exascale computing infrastructures.

The specific goal of NUMEXAS is the development of numerical methods for multiphysics problems in engineering based on validated models that enable scaling to millions of cores along the complete simulation pipeline.

The major challenge in NUMEXAS will be the development of a new set of numerical methods and computer codes that will allow industries, governments and academia to routinely solve multidisciplinary large-scale class problems in applied sciences and engineering with high efficiency and simplicity. We strive to demonstrate good scalability of up to several tens of thousands of cores in practice and to predict the theoretical capability of significant further performance gains with even higher orders of numbers of cores.

The NUMEXAS methods and codes will be the main project outcomes that will be disseminated and exploited by the partners. Emphasis will be put on the dissemination and exploitation of the NUMEXAS outputs among SMEs in Europe.

In order to achieve the above mentioned goals, improvements are required along the whole simulation pipeline, including parallel pre-processing of analysis data and mesh generation, parallel, scalable, parallel field solvers in fluid, solid mechanics and coupled problems, optimum design parallel solvers considering uncertainties and parallel post-processing of numerical results.

5.4.1.8. *H4H*

The H4H project is follow-on to the highly successful ParMA ITEA 2 project. The objective of H4H is to provide compute-intensive application developers with a highly efficient hybrid programming environment for heterogeneous computing clusters composed of a mix of classical processors and hardware accelerators. The project was funded in 2010 by the following agencies: DGCIS (Ministry of Industry, France), Federal Ministry of Education and Research (Germany) and Ministry of Industry, Energy and Tourism (Spain).

H4H leveraged and consistently advanced the state-of-the-art in several key software areas: programming models and associated runtimes, performance measurement and correctness tools, smart translation, intelligent mapping of processes/threads to hardware topology and resources, dynamic automatic tuning, and prediction of the execution time of a parallel application on different platforms:

- MAQAO – Modular Assembly Code Quality Analyzer and Optimize, developed by UVSQ
- PAS2P – a toolset to automatically extract the most significant behaviour (phases) of parallel applications, the Parallel Application Signature, that by its execution on different parallel computers, lets us predict the applications' performance. PAS2P was developed by UAB-CAOS
- Scalasca – a toolset to analyse the performance behaviour of parallel applications and to identify opportunities for optimization. Scalasca was developed by JSC
- STEP – Transformation System Parallel Execution. It is a tool transforming OpenMP programs into MPI programs. It is included in a compilation framework providing array region analysis, developed by Télécom SudParis (TSP).
- ThreadSpotter – performs automatic analyses of the whereabouts of a binary and suggest hands-on changes at the source level to a programmer. ThreadSpotter was developed by Rogue Wave Software.
- Valgrind – a public instrumentation framework for building dynamic analysis tools. The distribution currently includes: a memory error detector; two thread error

detectors; a cache and branch prediction profiler; a call-graph generating cache profiler; and a heap profiler.

- VampirTrace – a run-time library and toolset for instrumentation and tracing of software applications using Open Trace Format (OTF). The traces can be visualized by the Vampir and Scalasca tools. VampirTrace was developed by Zentrum für Informationsdienste und Hochleistungsrechnen (ZIH) at Technische Universität Dresden.

5.5 Chapter Summary

In this chapter, we described the activities related to exascale systems, both on computing and data management levels. Data management was described in terms of delivering big input data, movement between sites and the performance related to extremely fast external cache memories used for intermediate steps and final results of computations. The hardware solutions of computing elements are Intel, OpenPOWER, but products from companies trying to deliver unconventional technologies such as Adapteva Epiphany and Kalray MPPA are also included. The major goal here is to deliver fast enough, scalable but energy efficient computing nodes.

Existing storage solutions that were mostly designed in the terascale era will not scale to sustain I/O bandwidth necessary to support exascale machines. Although no clear solution has been identified yet, experts seem to agree on several aspects. First of all, the existing POSIX semantics should be abandoned or loosened, as they will not scale to the required number of requests per second. One option could be to move towards object based storage solutions with loose requirements for consistency. Furthermore, a more layered architecture is necessary with very fast intermediate storage, probably in the form of NVRAM solutions. Although SSD based storage is becoming cheaper it is predicted that it will still not be economically feasible as sole storage solution for hundreds of petabytes of data produced by exascale machines, thus a hybrid storage solution based on flash and HDD storage will be necessary.

In the next few years NAND storage devices might replace traditional HDDs. Nowadays the cost of building a system with the same performance on HDD and NAND is comparable, but on HDD with the same GBps/€ we receive ten times more capacity. The technology of NAND devices is still being improved; however the writing of NAND cells is a destructive process. High-end devices allow writing their whole capacity 30 times a day for a 5 years timeframe (DWPD), which makes these devices usable for Big Data and HPC.

The cost of storage is usually not bigger than 5–10% of the total cost of a supercomputer. When it is chosen well, it should work for more than one generation of supercomputers installed at the HPC site.

There are two ways to build reliable storage systems for Big Data and HPC:

1. File system with build-in data protection running on inexpensive hardware. All file systems with data protection are commercial. Free of charge file systems only promise to have this feature in the future.
2. File system without data protection (i.e. Lustre) running on reliable and expensive hardware.

The third combination of inexpensive hardware and file system without data protection is not recommended due to the high number of disks needed to build a file system ready for big data and due to the high probability of hardware failures.

We can observe in the last years a joint effort of HPC centres, R&D institutions, end users and HPC industry to propose new solutions that will bring us closer to exascale computing.

The solutions are on several levels: hardware technologies, new algorithms, programming environments and system software stacks (Table 7).

	Hardware develop-ment	Data manage ment	Programming environments	Communi cation libraries	Apps optimisation	New algorithms
DEEP	X		X		X	
DEEP-ER	X	X	X	X	X	
Mont-Blanc Mont-Blanc 2	X		X		X	
CRESTA			X	X	X	X
EPiGRAM			X	X	X	
EXA2CT			X		X	X
NUMEXAS			X	X	X	X
H4H			X	X	X	X

Table 7: Major topics covered by the exascale projects

6 PRACE and the European HPC Ecosystem in a Global Context

PRACE does not exist in a vacuum, but interacts with other organisations both in Europe and the rest of the world. This section takes a look at how it fits into the wider HPC ecosystem.

6.1 A global HPC policy in Europe

In February 2012 The European Commission published a communication that underlines the strategic nature of HPC[5]. 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[6]. The conclusions were a clear recognition of the need for an EU-level policy in HPC addressing the entire HPC ecosystem. PRACE and the European Technology Platform for High Performance Computing (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.

This policy has now been definitely adopted and its implementation has started, relying on two tightly related aspects and instruments:

[5] "High Performance Computing: Europe's place in a Global Race" - Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the regions, 15.2.2012, COM(2012)15

[6] 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

- A contractual Public-Private Partnership (cPPP) for HPC has been signed December 17th, 2013, by the ETP4HPC and the EC
- December 11th, 2013, Horizon 2020 has opened its first calls in the context of the Work Programme 2014-215, with a significant amount of funding for HPC during this first period, for HPC technologies as well as infrastructures and Centres of Excellence.

These two aspects are further described and commented in the next two sections.

6.2 ETP4HPC and HPC cPPP

An industry-led forum, which also has many members from the HPC research community, ETP4HPC[7] is the answer to the Commission to the need for a strong HPC technology pillar (supply side) in Europe – "competitive European HPC technologies for Europe science and industry competiveness" summarizes ETP4HPC credo and objectives. In 2013 ETP4HPC has been formally established by the European Commission as one of the recognized European Technology Platforms (ETPs). ETP4HPC is now included in the list annexed to the strategy for European Technology Platforms - ETP 2020[8]. This made ETP4HPC a distinguished voice for the definition of European HPC priorities and related R&D&I programmes. As of mid-2014, ETP4HPC has regularly grown, from 16 founding members in 2012, to reach 56 members – a mix of companies and research organization - out of which 23 are SMEs, with a mix of ISVs, services providers, in addition to integrators or hardware companies.

ETP4HPC had previously released its Strategic Research Agenda in April 2013[9]. 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 two 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 are indeed aspects of paramount importance, in addition to exascale and beyond. This is because 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.

Brussels, 17th of December 2013: ETP4HPC signed an Agreement with the European Commission (EC) to form a contractual Public-Private Partnership (cPPP) for the development of a European HPC eco-system. The HPC cPPP is one of the five new PPP's created within the Horizon 2020 Programme in addition to the four PPP's already operating in the Commission's previous programmes[10].

[7] http://www.etp4hpc.eu/
[8] "Individual ETPs," [Online]. Available: http://cordis.europa.eu/technology-platforms/individual_en.html.
[9] "ETP4HPC Strategic Research Agenda - Achieving HPC leadership in Europe", April 2013, http://www.etp4hpc.eu/wp-content/uploads/2013/06/ETP4HPC_book_singlePage.pdf
[10] http://europa.eu/rapid/press-release_MEMO-13-1159_en.pdf

This Agreement paves the way for a structured dialogue between the Public side (the EC) and the Private side (the ETP) in order to facilitate the achievement of research and investment objectives in the area of HPC technology provision, application development and sustainable HPC infrastructure. The EC's financial contribution to this research programme is 700 million Euros (142 million in 2014-2015) and for each Euro invested by the Public side, the Private side is expected to generate another four Euro's worth of research, innovation, investment and commercialisation. The ultimate objective of this research programme is to increase Europe's competitiveness, create jobs, stimulate innovation and build a world-class HPC value chain.

Formally, the cPPP has been signed by ETP4HPC only while reserving seats in its Board for the forthcoming Centres of Excellence for Computing Applications (see Figure 29, taken from [26]). PRACE is sitting aside the cPPP and has established a constructive dialogue with ETP4HPC so that their responsibilities are clearly divided reflecting their respective domains and expertise, and that they combine their efforts to strengthen Europe's place on the global HPC stage.

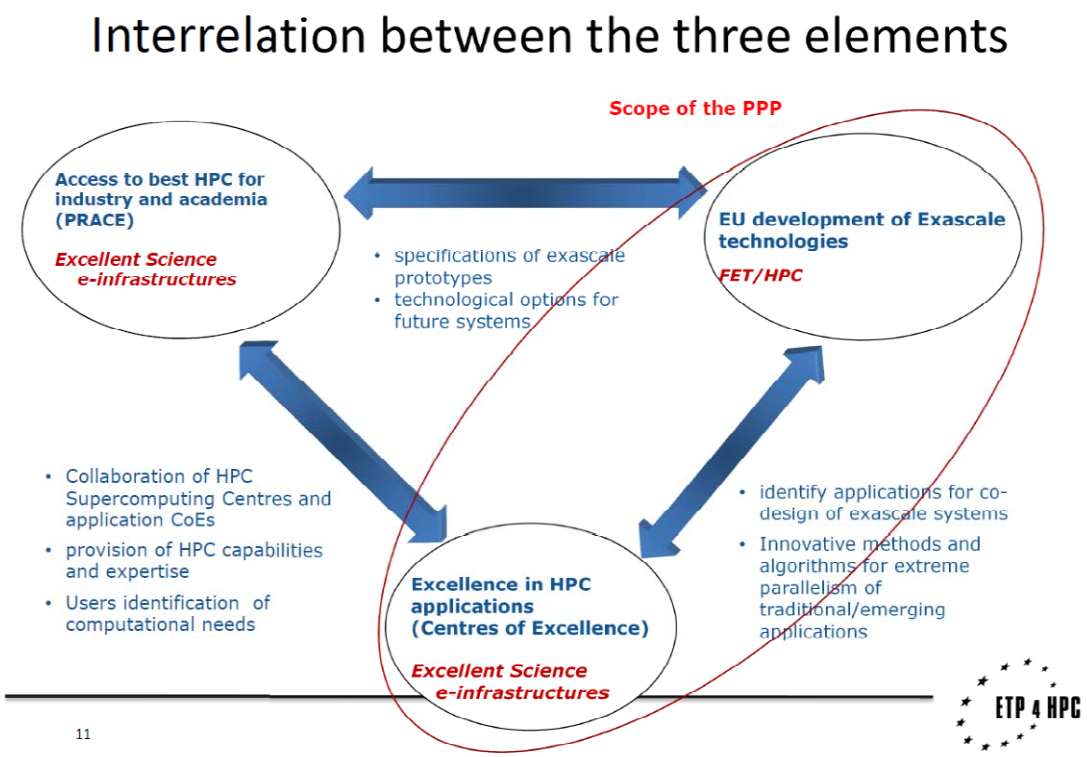


Figure 29: HPC cPPP scope and interaction between technologies (ETP4HPC scope), infrastructures (PRACE scope) and applications (CoEs)

April 9th, 2014, an HPC Info Day was organized by ETP4HPC and the EC[11]. This Info Day was the first public event within the charter of the PPP on HPC. The programme of this Info Day included the PPP's objectives, the European HPC strategy, PRACE contribution to the ecosystem, and the Calls for Proposals related to HPC in the Work Programme 2014-2015 of the Excellent Science pillar of Horizon 2020 (Future and Emerging Technologies (FET) and e-infrastructures).

The event gathered 140 participants from 20 countries at Institut de Physique du Globe de Paris. It offered an opportunity for many direct questions and exchanges with DG-CONNECT – as well as networking between participants.

[11] http://www.etp4hpc.eu/news/hpc-public-private-partnership-info-day-april-9-2014-paris/

6.3 H2020 calls: technologies, infrastructures, centres of excellence

As mentioned above, a first series of calls relating to HPC has been put in place in Horizon 2020, as soon as December 11th (opening of Work Programme 2014-2015, first period of H2020). The EC's financial contribution to this research programme is 142 million in 2014-2015. The cPPP provides a framework for a structured vision and orientation of this programme in the longer term, with perspectives of a sustained prioritization and further funding all along H2020 lifespan. The HPC programme is under the umbrella of Pillar 1 "Excellence in Science" of Horizon 2020 and managed by Directorate C of DG-CONNECT[12].

An overview of these calls can be found in Figure 30, taken from [26].

HPC related Calls WP 2014-2015

	2014 EUR million	2015 EUR million	Call Deadline
EINFRA-4-2014 - Pan-European HPC infrastructure and services	15		02/09/2014 - 17:00 Brussels time
EINFRA-5-2015 - Centres of Excellence (CoE) for computing applications		*40 (tbc)*	2015 *(date tbc)*
EINFRA-6-2014 - Network of HPC Competence Centres for SMEs	2		02/09/2014 - 17:00 Brussels time
FETHPC1-2014 HPC Core Technologies, Programming Environments and Algorithms for Extreme Parallelism and Extreme Data Applications	93,4		25/11/2014 at 17.00.00 Brussels time
FETHPC 2 - 2014: HPC Ecosystem Development	4		25/11/2014 at 17.00.00 Brussels time

Figure 30: HPC related calls in Work Programme 2014-2015 of Horizon 2020

Regarding technologies, a specific programme FETHPC-2014 "TOWARDS EXASCALE HIGH PERFORMANCE COMPUTING" has been embedded in Future and Emerging Technologies (FET pro-active) branch of H2020 Excellence in Science[13]. Paving the way to exascale technological capacity, this call mostly borrows, for its FETHPC1 part with funding for R&I projects, from ETP4HPC SRA recommendations and emphasizes four subtopics:

- HPC core technologies and architectures
- Programming methodologies, environments, languages and tools
- APIs and system software for future extreme scale systems
- New mathematical and algorithmic approaches for existing or emerging applications on extreme scale systems

Needless to say, the call is open to proposals from all organisations, whether or not they are involved in ETP4HPC or the cPPP.

A specific Coordination and Support Action call (FETHPC2) has been established as well, actually twofold:

[12] http://ec.europa.eu/dgs/connect/en/content/einfrastructures-computational-infrastructure
[13] http://ec.europa.eu/research/participants/portal/desktop/en/opportunities/h2020/calls/h2020-fethpc-2014.html

- "Coordination of the HPC strategy: The aim is to support the implementation of a common European HPC strategy through coordination of the activities of stakeholders such as the European Technology Platform for HPC (ETP4HPC), PRACE, application owners and users (including emerging HPC applications), the European exascale computing research community, the open source HPC community, related activities in other parts of H2020, etc."
- "Excellence in High Performance Computing Systems: The aim is to boost European research excellence on the key challenges towards the next generations of high-performance computing systems (such as energy efficiency, complexity, dependability and cutting across all levels – hardware, architectures, programming, applications)"

There are, in Horizon 2020, a number of other HPC-related activities, which can be found either in Pillar 2 – Industrial Leadership, or Pillar 3 – Societal Challenges, but the bulk of HPC in the cPPP sense is consistently gathered in Pillar 1. For instance FETHPC1 clearly states that "… This activity [FETHPC1] will be coordinated with complementary work in LEIT/Advanced Computing, LEIT/Photonics, and ECSEL (Electronic Components and Systems for European Leadership) that will develop basic system technology that is relevant to the needs of exascale computing (e.g. microprocessors, photonics components, interconnects or system software, programming environments for critical/real time systems, etc.)..."

Indeed, in LEIT/Advanced Computing an emphasis on low power computing can be found, with wider perspective than, but possible synergy with, HPC – wider in the market/niche sense, i.e. micro-servers with less computing density but more market volume than high-end HPC servers.

The other calls that pertain to HPC in the cPPP broader sense are found in EINFRA calls (H2020-EINFRA-2014/2015):

- Centres of Excellence for computing applications[14]
  "Establishing a limited number of Centres of Excellence (CoE) is necessary to ensure EU competitiveness in the application of HPC for addressing scientific, industrial or societal challenges. CoEs will be user-focused, develop a culture of excellence, both scientific and industrial, placing computational science and the harnessing of "big data" at the centre of scientific discovery and industrial competitiveness. CoEs may be "thematic", addressing specific application domains such as medicine, life science or energy; "transversal" on computational science (e.g. algorithms, analytics, numerical methods etc.); or "challenge-driven", addressing societal or industrial challenges (e.g. ageing, climate change, clean transport etc.); or a combination of these types. This topic will be carried out in the context of the Public-Private Partnership (PPP) in HPC, contributing to the implementation of the EU strategy on High Performance Computing (HPC), in particular to achieving excellence in HPC application delivery and use."
- Network of HPC Competence Centres for SMEs[15]
  "HPC competence centres have been set up in some Member States to facilitate access and take-up by industry and in particular SMEs of HPC services. As yet these centres do not cover the whole of Europe. Supporting one network of HPC competence centres will promote access to computational expertise anywhere in Europe and enable the dissemination of best practice in HPC industrial use particularly for SMEs. This topic contributes to the implementation of the European HPC strategy, in particular to foster the use of HPC by SMEs."

[14] http://ec.europa.eu/research/participants/portal/desktop/en/opportunities/h2020/topics/2143-einfra-5-2015.html
[15] http://ec.europa.eu/research/participants/portal/desktop/en/opportunities/h2020/topics/2140-einfra-6-2014.html

- Pan-European High Performance Computing infrastructure and services[16]
  "In order to create a world-class pan-European infrastructure, and to provide state-of-the-art services and access to this infrastructure to users, independently of location, the HPC resources in Europe need to be further pooled, integrated and rationalised. This topic contributes to the implementation of the EU strategy on High Performance Computing (HPC), in particular by providing access to the best supercomputing facilities and services for both industry and academia, and complements the activities of the Public-Private Partnership (PPP) in HPC in order to implement the HPC strategy."
  This latter call is clearly a place for a PRACE next Implementation Phase project submission, which is actually under construction under "PRACE 4IP" name.

6.4 PRACE: resources, usages w.r.t. other continents

PRACE has regularly been 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 ca. 15 PFlop/s, a significant fraction of which is reserved for PRACE:

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

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 2014 Top500 list[17], all in the Top50, resp. at ranks 8, 12, 17, 26, 41 and 44 (so, 1 of these systems is in the Top10 and 3 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. PRACE has not gone strongly "hybrid" yet either, unlike for instance a significant fraction of the Top10 systems (Tianhe-2@NUDT, Titan@ORNL, PizDaint@CSCS, STAMPEDE@TACC) or BlueWaters@NCSA (not in the Top500 but comparable to Top10 systems).

Since mid-2010 PRACE has been maintaining a steady growth from 363 to the order of 1200 million core hours granted every six months through Regular Calls – including ca. 655 million of x86 core hours and ca. 525 million of BG/Q core hours (Power BQC 16C 1.6GHz) in the Regular Call 9 of PRACE, granted July 2014 for allocations starting Fall of 2014 – for one year.

PRACE has actually reached an operational plateau, all the more approaching the end of its initial period – work is in progress regarding an updated model for a second PRACE period, after 2015, and the renewal of the systems. The plateau is perceptible also through the

[16] http://ec.europa.eu/research/participants/portal/desktop/en/opportunities/h2020/topics/2139-einfra-4-2014.html
[17] http://www.top500.org/lists/2014/06/

increasing demand during PRACE bi-annual "Regular calls", since the offer is steady and more and more users have matured their codes and capacity of going petascale or near-petascale – in short, being able to credibly apply for Tier-0 resources. So the "selection pressure" is increasing.

6.4.1 *A quick comparison with US and Japan allocation systems*

6.4.1.1. *INCITE in the USA*

INCITE[18] is a programme with many similarities to PRACE, although not strictly comparable in terms of frequency and duration of calls and allocations:

- Best science, capability – criteria of high-impact, computationally intensive research campaigns in a broad array of science, engineering, and computer science domains
- Annual call for proposals of individuals and teams of researchers from academia, national laboratories, and industry
- Awards of one, two, or three years are granted
- INCITE has been boosted by recent multi-petascale systems deployment, hybrid or not (MIRA, TITAN)

Some indications on current and previous calls (quotations from https://proposals.doeleadershipcomputing.org/allocations/calls/incite2015):

- INCITE is currently soliciting proposals of research for awards of time on the 27-petaflop Cray XK7, "Titan", and the 10-petaflop IBM Blue Gene/Q, "Mira", beginning calendar year (CY) 2015. Nearly six billion core-hours will be allocated for CY 2015. Average awards per project for CY 2015 are expected to be on the order of 75 million core-hours for Titan and 100 million core-hours for Mira, but could be as much as several hundred million core hours.
- Last year's call for proposals resulted in 59 projects (38 new, 21 renewals) awarded 5.8 billion core-hours for CY 2014. The acceptance rate for new proposals was 36 percent. Representative awards include the following
  - **Biophysics** (75 million core-hours) "Assembling and sustaining the 'acid mantle' of the human skin barrier"
  - **Accelerators** (50 million core-hours) "Intensity-dependent Dynamics in Fermilab and CERN Accelerators"
  - **Engineering** (100 million core-hours) "Combustion stability in complex engineering flows"
  - **Climate** (150 million core-hours) "High Resolution Simulation for Climate Means, Variability and Extreme"
  - **Materials Science** (60 million core-hours) "Innovative Simulations of High-Temperature Superconductors"
  - **Plasma Physics** (239 million core-hours) "High-fidelity simulation of tokamak edge plasma transport"
  - **Computer Science** (40 million core-hours) "Collaborative Research into Exascale Systemware, Tools and Applications"

INCITE proposals are accepted between mid-April and the end of June. We can thus roughly estimate INCITE to be 2 to 2,5 times bigger than PRACE if we compare on one running year (1 INCITE call vs. 2 PRACE calls). This could be further weighted by comparing the respective fractions of GPU, x86 or BQC processor cycles in the bouquet of core*hours.

[18] http://www.doeleadershipcomputing.org/incite-PROGRAM/

6.4.1.2. *RIKEN/HPCI system around K computer in Japan*

K computer and other major supercomputers in Japan are connected via high-speed networks to form the High Performance Computing Infrastructure[19] for a total computing power of 25 PFlop/s – see Figure 31 below. The size of resources and granting processes also exhibit similarities with PRACE, from application to reporting of results.

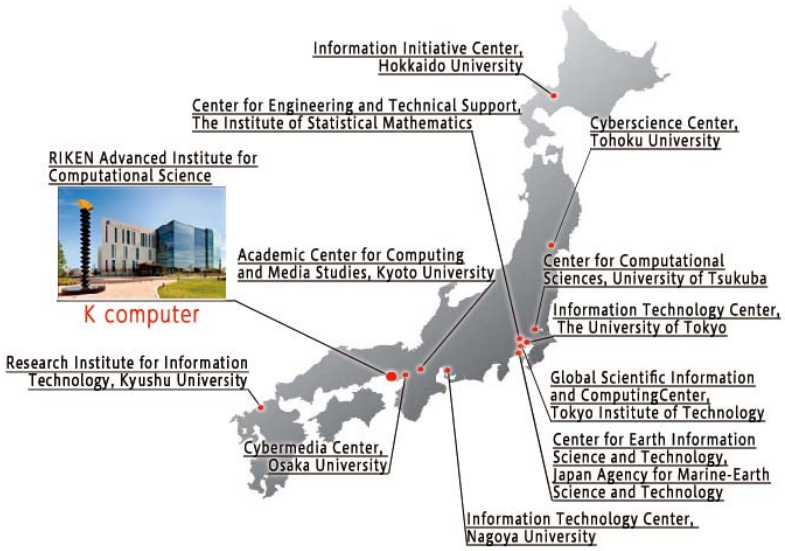


Figure 31: Providers of the computational resources that constitute the HPCI system in Japan

HPCI Operation Office promotes the utilization of the HPCI by selecting the users and receiving proposals for the use of entire HPCI within a framework of "Innovative High Performance Computing Infrastructure Project" set up by the MEXT (Ministry of Education, Culture, Sports, Science and Technology in Japan). Single Sign On (SSO) is available in the HPCI.

In order to utilize the K computer or computers other than the K computer provided through the HPCI System, one has to apply for the call for the proposals normally called once or twice a year, and let the proposals go through a screening process before possible grant and allocation start. Screening of project proposals is carried out by the Project Screening Committee which consists of experts in industry and academia. This Committee will evaluate the project proposals including the application form submitted by the applicants, based on published guidelines.

The computational resources of the K computer are made available to the users of several categories as shown in Figure 32. The resource to be used by the users of the General Use Category which represents about 33% of the total resource includes the resources for the Industrial Use Projects and Junior Researcher Promotion Projects.

[19] https://www.hpci-office.jp/folders/e_guide
https://www.hpci-office.jp/pages/e_concept
http://www.icri2014.eu/sites/default/files/presentations/Satoshi MATSUOKA.pdf

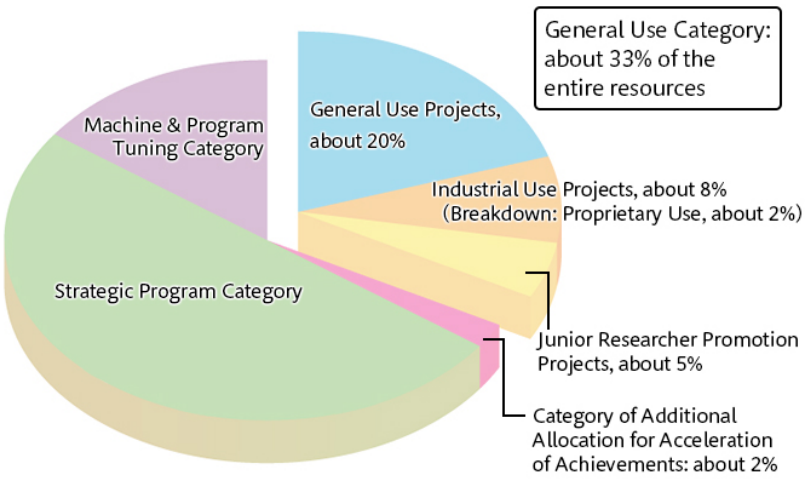


Figure 32: Available resources of the K computer and fractions of the kinds of allocations

In addition to this category, there are categories such as the Strategic Program Category which are not provided through public calls for proposals. The awarding of the projects in the Strategic Program Category will be strategically decided by the Japanese Government. What is comparable to PRACE is the ca. 1/3 of the resources in General User Category of the K system, plus analog access on other computers (however a part of this latter set of resources may probably be more comparable to PRACE DECI Tier-1 allocations).

7 Conclusion and Summary

As can be seen in the analysis of the latest developments on the Top500 list, the HPC industry seems to be on a plateau currently. Systems are getting older and the replacement rate is historically low. At the same time a wide range of approaches to reach the exascale target are being tried out both by research groups and vendors. In the coming years we will see these products being tested in the market, on large petascale systems at first. Which architectures will dominate in the future is an open question, but it is quite possible that we will see more heterogeneous architectures than today.

Water cooling is currently the leading candidate for handling the heat generated by the large systems, but more work stills needs to be done to answer the question of what to do with the heated water. Using it to heat buildings is one approach, but needs infrastructure nearby to distribute the heat. Feeding it back into adsorption chillers is another promising technique.

Optimizing applications will be a growing concern in the future. But in order to enable application optimization, measurements at different levels are needed. A programmer may currently be mainly concerned with CPU timing and communication latency, but to do a system wide optimization care needs to be taken to ensure power efficiency. The rising cost of electricity will give incentives for scheduling the jobs requiring the most power at times when the spot price for electricity is low. Variations in the power consumption of future systems will lead to a need for more communication with the company providing the power grid.

The infrastructure workshops have been a good example for bringing together people working on site infrastructure.

European collaboration will be needed for developing exascale technologies, and not relying totally on other countries. Currently US companies are developing most of this technology, and the European effort is mostly system integration and software development. Through the HPC-related efforts in the H2020 programme, Europe will possibly be able to change the status quo in future.

8 Annex

8.1 Infrastructure workshop program

5th European workshop on HPC centre infrastructures

1st to 3rd April 2014

Domaine de Frémigny - Bouray-sur-Juine

Très Grand Centre de Calcul du CEA

Bruyères-le-Châtel

Day 1	Tuesday, April 1st
08:45 - 09:00	Welcome and introduction –J.P. Nominé
09:00 - 09:45	William Kramer – NCSA, USA
09:45 - 10:15	EE HPC WG - Natalie Bates
10:15 - 10:45	Allan Williams – NCI, Australia
10:45 - 11:30	Break
11:30- 12:00	John Dolphin – AWE, UK
12:00 - 12:30	Alain Beuraud – Meteo France
12:30 – 13:00	Peter Marksteiner– Vienna Scientific Cluster - HPC in oil: a large cluster with liquid immersion cooling
13:00 – 14:00	Lunch
14:15 - 16:15	Direct liquid cooling panel : large deployments and vendors perspectives Moderators : Norbert Meyer, Gert Svensson Panelists: • Laurent Cargemel, Bull • Wilfried Oed, Cray • Nicolas Dubé, HP • Rudiger Wolff, SGI
16:15 - 16:30	CEA site update and introduction to the visit - François Robin
16:30	Break
17:00 - 19:30	TGCC visit @ CEA
Day 2	Wednesday, April 2nd
08:30 - 09:00	Guillermo Aguirre –BSC, Spain
09:00 - 09:30	Mike Patterson - Intel
09:30 - 10:00	Alban Schmutz – OVH, France
10:00 - 10:30	Steven Hammond – NREL, USA
10:30 – 11:00	Break
11:00 - 11:30	Mattias Ganslandt – Lund University, Sweden
11:30 - 12:00	Herbert Huber – LRZ, Germany
12:00 - 12:30	Ladina Gilly – CSCS, Switzerland
12:45 – 14:00	Lunch
14:00 - 16:30	European HPC R&D towards energy-efficient exascale systems Moderators : Guillermo Aguirre, Norbert Meyer Panelists: • Jochen Kreutz, FJZ : DEEP • Petar Radojkovic, BSC : MontBlanc • CINECA– Carlo Cavazzoni; Eurotech – Giovanbattista Mattiussi : EURORA • Torsten Wilde- LRZ ; Thomas Blum - Megware
Day 3	Thursday, April 3rd
08:30 – 12 :15	PRACE closed session