E-Infrastructures
H2020-EINFRA-2014-2015

EINFRA-4-2014: Pan-European High Performance Computing Infrastructure and Services

PRACE-4IP

PRACE Fourth Implementation Phase Project

Grant Agreement Number: EINFRA-653838

D5.1
Market and Technology Watch Report Year 1

Final

Version:	1.1
Author(s):	Jean-Philippe Nominé, CEA
Date:	19.04.2016

Project and Deliverable Information Sheet

PRACE Project	**Project Ref. №: EINFRA-653838**	
	Project Title: PRACE Fourth Implementation Phase Project	
	Project Web Site: http://www.prace-project.eu	
	Deliverable ID: D5.1	
	Deliverable Nature: Report	
	Dissemination Level: Public	**Contractual Date of Delivery:** 30 / 04 / 2016
		Actual Date of Delivery: 30 / 04/ 2016
	EC Project Officer: Leonardo Flores Añover	

Document Control Sheet

Document	**Title: Market and Technology Watch Report Year 1**	
	ID: D5.1	
	Version: 1.1	**Status: Final**
	Available at: http://www.prace-project.eu	
	Software Tool: Microsoft Word 2010	
	File(s): D5.1.docx	
Authorship	**Written by:**	Jean-Philippe Nominé, CEA
	Contributors:	Felip Moll, BSC Oscar Yerpes, BSC Giannis Koutsou, CASTORC Francois Robin, CEA Carlo Cavazzoni, CINECA Olli-Pekka Lehto, CSC Dirk Pleiter, FZJ Michael Stephan, FZJ Eric Boyer, GENCI Philippe Segers, GENCI Ioannis Liabotis, GRNET Dimitrios Dellis, GRNET Nikolaos Nikoloutsakos, GRNET Branislav Jansik, IT4I-VSB Filip Stanek, IT4I-VSB Gert Svensson, KTH Andreas Johansson, LiU Michael Ott, BADW-LRZ Torsten Wilde, LRZ Radek Januszewski, PSNC Norbert Meyer, PSNC Huub Stoffers, SURFsara
	Reviewed by:	Hank Nussbacher, IUCC Florian Berberich, JUELICH - PMO
	Approved by:	MB/TB

Document Status Sheet

Version	Date	Status	Comments
0.1	29/12/2016	Draft outline	Detailed outline as agreed during Dec. 10, 2016, F2F meeting
0.2	25/03/2016	Partial draft	Inclusion of contributions received until March 20th
0.3	29/03/2016	Draft	First full inclusion of almost all contributions (still missing a couple)
0.4	01/04/2016	Draft	Misc. addenda (sections 2.3, 3.4, 4.1, chapter 6, draft conclusion)
0.5	04/04/2016	Draft	Almost full version for finalisation telco of April 4th
0.6	04/04/2016	Draft	Editorial improvements of v0.5
0.7	05/04/2016	Draft	Various corrections
1.0	06/04/2916	Final	For internal PRACE review
1.1	19/04/2016	Final version	Corrected for MB/TP approval, from internal reviews

Document Keywords

Keywords:	PRACE, HPC, Research Infrastructure, Market Survey, Technology Watch

Disclaimer

This deliverable has been prepared by the responsible Work Package of the Project in accordance with the Consortium Agreement and the Grant Agreement n° EINFRA-653838. It solely reflects the opinion of the parties to such agreements on a collective basis in the context of the Project and to the extent foreseen in such agreements. Please note that even though all participants to the Project are members of PRACE AISBL, this deliverable has not been approved by the Council of PRACE AISBL and therefore does not emanate from it nor should it be considered to reflect PRACE AISBL’s individual opinion.

Copyright notices

Table of Contents

List of Figures

List of Tables

References and Applicable Documents

[1] PRACE: http://www.prace-project.eu

[2] Top 500: www.top500.org/

[3] Green 5000: http://www.green500.org/

[4] HPCG Benchmark: http://hpcg-benchmark.org/

[5] HPL Benchmark: http://www.top500.org/project/linpack/

[6] http://www.hpcwire.com/2015/11/18/sugon-vp-on-global-market-strategy-the-vmware-venture-and-robotic-immersive-cooling/

[7] http://www.hpcwire.com/2016/03/11/idc-server-report-china-leads-growth-ibm-power-strengthens-arm-stumbles/

[8] http://www.idc.com/getdoc.jsp?containerId=prUS41076116

[9] http://www.marketsandmarkets.com/PressReleases/Quantum-High-Performance-Computing.asp

[10] http://hpcadvisorycouncil.com/events/2016/stanford-workshop/pdf/Snell.2016TrendsHPCandHyperscale.Intersect360Rsrch.pdf

[11] http://www.marketsandmarkets.com/Market-Reports/Quantum-High-Performance-Computing-Market-631.html?gclid=CNDg0cmf5ssCFZcy0wodeT8OVw

[12] https://www.whitehouse.gov/the-press-office/2015/07/29/executive-order-creating-national-strategic-computing-initiative

[13] http://energy.gov/articles/department-energy-awards-425-million-next-generation-supercomputing-technologies

[14] http://www.hpcwire.com/2016/02/10/final-obama-budget-fy17/

[15] https://asc.llnl.gov/CORAL/

[16] http://www.theregister.co.uk/2015/04/10/us_intel_china_ban/

[17] http://www.scientific-computing.com/features/feature.php?feature_id=467

[18] IDC Report - SMART 2014/0021 – “High Performance Computing in the EU: Progress on the Implementation of the European HPC Strategy” http://ec.europa.eu/newsroom/dae/document.cfm?doc_id=10334

[19] http://www.hpcwire.com/off-the-wire/riken-selects-fujitsu-develop-new-supercomputer/

[20] http://www.aics.riken.jp/en/postk/outcome

[21] Japan Concludes Exascale Feasibility Study: http://www.hpcwire.com/2014/12/03/japan-concludes-exascale-feasibility-study/

[22] Japan Preps for HPC-Big Data Convergence: http://www.hpcwire.com/2015/06/18/japan-preps-for-hpc-big-data-convergence/

[23] https://ec.europa.eu/programmes/horizon2020/en/h2020-section/high-performance-computing-hpc

[24] http://ec.europa.eu/research/participants/portal/desktop/en/opportunities/h2020/topics/2119-einfra-11-2016.html

[25] http://ec.europa.eu/research/participants/portal/desktop/en/opportunities/h2020/calls/h2020-fethpc-2016-2017.html

[26] http://ec.europa.eu/research/participants/data/ref/h2020/wp/2016_2017/main/h2020-wp1617-fet_en.pdf

[27] SRA http://www.etp4hpc.eu/en/sra.html

[28] ETP4HPC http://www.etp4hpc.eu/

[29] ETP4HPC members http://www.etp4hpc.eu/en/membership.html

[30] EXDCI https://exdci.eu/

[31] https://exdci.eu/events/european-hpc-summit-week

[32] http://www.etp4hpc.eu/en/news/25-exdci-workshop-in-rome-september9-and0.html

[33] https://ec.europa.eu/digital-single-market/en/news/study-high-performance-computing-eu-progress-implementation-european-hpc-strategy-final-report

[34] https://ec.europa.eu/digital-single-market/en/news/communication-high-performance-computing-hpc-europes-place-global-race

[35] https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud

[36] http://knowledgebase.e-irg.eu/documents/243153/299805/IPCEI-HPC-BDA.pdf

[37] http://www.ccs.tsukuba.ac.jp/eng/wordpress/wp-content/uploads/2014/04/CCS-MitsuhisaSato.pdf

[38] Tom R. Halfhill, "POWER8 hits the merchant market," The Linley Group, Microprocessor Report, December 2014.

[39] E. J. Fluhr et al., "POWER8: A 12-core server-class processor in 22nm SOI with 7.6Tb/s off-chip bandwidth," ISSCC 2014, doi:10.1109/ISSCC.2014.6757353.

[40] http://openpowerfoundation.org/ (accessed 19.03.2016)

[41] https://www-03.ibm.com/press/us/en/pressrelease/42980.wss (accessed 19.03.2016)

[42] http://openpowerfoundation.org/wp-content/uploads/2015/03/Ashley-John_OPFS2015_NVIDIA_031215.pdf (accessed 19.03.2016)

[43] http://www.nallatech.com/wp-content/uploads/IBM_CAPI_Users_Guide.pdf (accessed 19.03.2016)

[44] http://openpowerfoundation.org/blogs/interconnect-your-future-mellanox-100gb-edr-capi-infiniband-and-interconnects/ (accessed 19.03.2016)

[45] http://on-demand.gputechconf.com/gtc/2015/presentation/S5786-Tjerk-Straatsma.pdf (accessed 19.03.2016)

[46] https://software.intel.com/en-us/articles/what-disclosures-has-intel-made-about-knights-landing

[47] http://www.bull.com/tera-1000-cea-completes-first-milestone-towards-exascale

[48] http://www.hardwareluxx.com/index.php/news/hardware/memory/36373-idf-2015-ddr4-roadmap-for-2015-2016.html (accessed 20.03.2016)

[49] https://www.skhynix.com/products.do?ct1=36&ct2=44&lang=eng (accessed 20.03.2016)

[50] http://www.hybridmemorycube.org (accessed 20.03.2016)

[51] Toshio Yoshida, "SPARC64TM Xifx: Fujitsu's Next Generation Processor for HPC," HotChips 2014, https://www.fujitsu.com/hk/Images/20140811hotchips26.pdf (accessed 20.03.2016)

[52] https://www.altera.com/en_US/pdfs/literature/ug/ug_hmcc.pdf (accessed 20.03.2016)

[53] https://www.micron.com/about/blogs/2016/february/gddr5x-has-arrived

[54] https://www.jedec.org/standards-documents/results/GDDR5X (access 26.03.2016)

[55] M. Webb, "Alternative Non-Volatile Memory Adoption Timeline," FlashMemory Summit 2015.

[56] R. Davis, "NVMe over Fabrics: Learning from early developments, " FlashMemory Summit 2015.

[57] http://www.ddn.com/products/infinite-memory-engine-ime14k/

[58] http://www.cray.com/products/storage/datawarp

[59] J. Handy, "Flash Technology: Annual Update," FlashMemory Summit 2015.

[60] J. Pappas, "Annual Update on Interfaces," FlashMemory Summit 2015.

[61] SNIA, "NVM Programming Model (NPM) Version 1.1," March 2015.

[62] D. Voigt, "Programming for Non-Volatile Memory," FlashMemory Summit 2015.

[63] JDEC, "High Bandwidth Memory (HBM) DRAM," JESD235A, November 2015

(http://www.jedec.org/standards-documents/docs/jesd235a
[64] http://top500.org/statistics/list/
[65] http://www.nextplatform.com/2015/11/16/intel-rounds-out-scalable-systems-with-omni-path/
[66] http://www.nextplatform.com/2015/09/14/arista-wields-tomahawks-in-25g-ethernet-price-war/
[67] Christelle Piechurski, BULL, BULL Booth, SC15, November 2015
[68] http://www.hoti.org/hoti23/slides/derraji.pdf
[69] Saïd Derradji, Thibaut Palfer-Sollier, Jean-Pierre Panziera, Axel Poudes, François Wellenreiter, The BXI Interconnect architecture, paper sent to the author
[70] Gilad Shainer, Mellanox, Conversation with Author, SC15, November 2015
[71] http://www.mellanox.com/page/products_dyn?product_family=227&mtag=switch_ib2_ic
[72] https://www.sgi.com/pdfs/4192.pdf
[73] Ulrich Krackhardt, Extoll, Extoll Booth, SC15, November 2015
[74] http://extoll.de/products/tourmalet
[75] Wilfried Oed, CRAY, Conversation with Author, SC15, November 2015
[76] http://www.cray.com/sites/default/files/resources/CrayXCNetwork.pdf
[77] http://www.fujitsu.com/global/Images/tofu-interconnect2_tcm100-1055326.pdf
[78] https://www.olcf.ornl.gov/summit/
[79] http://link.springer.com/article/10.1007%2Fs11704-014-3500-9#/page-1
[80] http://www.bull.com/
[81] http://atos.net/en-us/home/we-are/news/press-release/2015/pr-2015_11_12_01.html
[82] http://www.bull.com/sequana
[83] http://www.scientific-computing.com/news/news_story.php?news_id=2719
[84] Technical Advances in the SGI UV Architecture, 2012 https://www.sgi.com/pdfs/4192.pdf
[85] SGI product portfolio, https://www.sgi.com/products/servers/, retrieved March 2016
[86] Private correspondence between the Author and the SGI, 2015-2016.
[87] http://www.fujitsu.com/primergy (accessed 29.03.2016)
[88] http://docs.ts.fujitsu.com/dl.aspx?id=c0e740df-f6a3-4e5b-ae78-fd6567bade6c (accessed 29.03.2016)
[89] https://www.fujitsu.com/global/Images/key-hardware-technologies-for-the-next-generation-primehpc-post-fx10.pdf (accessed 29.03.2016)
[90] A. Caldeira et al., “IBM Power Systems S822LC Technical Overview and Introduction,” IBM Redpaper publication, 20
[91] http://www.openstack.org
[92] http://www.rdoproject.org
[93] https://wiki.openstack.org/wiki/Packstack
[94] https://wiki.openstack.org/wiki/Neutron/DVR.
[95] https://research.csc.fi/taito-supercluster
[96] http://www.docker.com
[97] https://www.ibm.com/developerworks/community/blogs/hpcgoulash/entry/a_whale_of_a_time?lang=en
[98] http://www.isc-events.com/isc15_ap/sessiondetails.htm?t=session&o=227&a=select
[99] http://investors.cray.com/phoenix.zhtml?c=98390&p=irol-newsArticle&ID=2112970
[100] https://www.nersc.gov/research-and-development/user-defined-images/
[101] http://www.lanl.gov/projects/apex/schedule.php
[102] Cacti® ,”The Complete RRDTool-based Graphing Solution”. 2004. 18 Jan. 2016 http://www.cacti.net/

[103] http://oss.oetiker.ch/rrdtool/
[104] Ovis, “Sandia National Laboratories”. 2007. 18 Jan. 2016 https://ovis.ca.sandia.gov/
[105] Birngruber, E., “Total recall: holistic metrics for broad systems performance” - 2015. http://dl.acm.org/citation.cfm?id=2835001
[106] “XDMoD Portal. 2011”. 18 Jan. 2016 <https://xdmod.ccr.buffalo.edu/>
[107] Forschungszentrum Jülich - JSC – “Components of LLview”. 2013. 18 Jan. 2016 http://www.fz-juelich.de/ias/jsc/EN/Expertise/Support/Software/LLview/llview-components_node.html
[108] “XALT” – Texas Advanced Computing Center. 2015. 18 Jan. 2016 https://www.tacc.utexas.edu/research-development/tacc-projects/xalt
[109] Rosales, C., “Remora: A Resource Monitoring Tool for Everyone…” - 2015. http://hust15.github.io/files/HUST15-Rosales-Remora-Slides.pdf
[110] “Redfish Resource Explorer”. 2015. 18 Jan. 2016 http://redfish.dmtf.org/
[111] PowerAPI: Sandia National Laboratories”. 2014. 18 Jan. 2016 http://powerapi.sandia.gov/
[112] Laros III, James H et al., “High Performance Computing-Power Application Programming Interface Specification Version 1.0”. - *Sandia National Laboratories, Tech. Rep. SAND2014-17061* (2014)
[113] https://collectd.org/
[114] http://graphite.wikidot.com/
[115] http://grafana.org/
[116] http://graphite.readthedocs.org/en/latest/tools.html
[117] https://www.elastic.co/products
[118] http://www.nersc.gov/research-and-development/apex/
[119] https://www.hepix.org/
[120] http://qnib.org/2015/10/29/melig-2/
[121] PRACE 2IP deliverable D5.2, “Updated Best Practices for HPC Procurement and Infrastructure” – August 2014
[122] http://exascale-projects.eu/ - an overview on FP7 funded projects
[123] http://www.deep-project.eu/deep-project/
[124] http://www.deep-er.eu
[125] http://www.scientificcomputing.com/news/2015/07/booster-system-installed-j%C3%BClich-completes-deep-supercomputer
[126] http://www.montblanc-project.eu/
[127] http://cresta-project.eu
[128] http://www.epigram-project.eu/
[129] https://projects.imec.be/exa2ct/
[130] http://www.numexas.eu/
[131] https://ec.europa.eu/programmes/horizon2020/en/news/21-new-h2020-high-performance-computing-projects
[132] http://www.etp4hpc.eu/pujades/files/European%20HPC%20Technology%20Handbook%20-%20SC15%20BOF.pdf
[133] http://www.etp4hpc.eu/en/euexascale.html
[134] https://ec.europa.eu/programmes/horizon2020/en/news/eight-new-centres-excellence-computing-applications
[135] https://exdci.eu/
[136] http://www.eurolab4hpc.eu/
[137] http://www.nesus.eu/
[138] https://itea3.org/all-projects/page-all.html
[139] http://ec.europa.eu/digital-agenda/en/pre-commercial-procurement
[140] https://www.humanbrainproject.eu/

[141] http://ec.europa.eu/research/participants/data/ref/h2020/other/wp/2016_2017/annexes/h2020-wp1617-annex-g-trl_en.pdf
[142] http://www.prace-ri.eu/ueabs/
[143] http://www.sage-project.eu/home.html
[144] https://www.tacc.utexas.edu/
[145] http://wccftech.com/nvidia-pascal-nvlink-200-gbs-interconnect-hbm2-stacked-memory-1-tbs-bandwidth-powering-hpc-2016

List of Acronyms and Abbreviations

aisbl	Association Internationale Sans But Lucratif (legal form of the PRACE-RI)
BXI	Bull eXascale Interconnect
CoE	Center of Excellence (for Computing Applications)
CORAL	Joint Collaboration of Oak Ridge, Argonne, and Lawrence Livermore US HPC Centers
CPU	Central Processing Unit
cPPP	contractual Public Private Partnership
CSA	Coordination and Support Actions (type of Horizon 2020 project)
CUDA	Compute Unified Device Architecture (NVIDIA)
DARPA	Defense Advanced Research Projects Agency
DP	Double Precision Floating point (usually in 64-bit)
DSM	Digital Single Market
EC	European Commission
EU	European Union
EFlop/s	Exaflop/s Exa (= 10^{18}) Floating point operations (usually in DP) per second, also EF/s
EXDCI	European Extreme Data & Computing Initiative
FETHPC	HPC programme of H2020 Future and Emerging Technologies branch
FP7	7th Framework Programme for Research and Technological Development of the European Union (Research and Innovation funding programme for 2007-2013.)
FGPA	Field Programmable Gate Array
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
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
GPU	Graphic Processing Unit
H2020	The EU Framework Programme for Research and Innovation 2014-2020
HBP	Human Brain Project
HPC	High Performance Computing; Computing at a high performance level at any given time; often used synonym with Supercomputing
HPCG	High Performance Conjugate Gradient – a benchmark developed as complement or alternative to HPL
HPDA	High Performance Data Analytics
HPL	High Performance LINPACK benchmark (used for Top500 ranking)
IPCEI	Important Project of Common European Interest
ISC	International Supercomputing Conference; European equivalent to the US based SCxx conference. Held annually in Germany.
KB	Kilo (= 2^{10} ~10^{3}) Bytes (= 8 bits), also KByte
KNC	Knights Corner, Intel® MIC Xeon® Phi™ processors (first generation)
KNF	Knights Ferry, Intel® MIC Xeon® Phi™ processors (prototype)
KNL	Knights Landing, Intel® MIC Xeon® Phi™ processors (second generation)
LINPACK	Software library for Linear Algebra
MB	Management Board (highest decision making body of the project)

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
MIC	Intel Many Integrated Core Processor Architecture
MPI	Message Passing Interface
NDA	Non-Disclosure Agreement. Typically signed between vendors and customers working together on products prior to their general availability or announcement.
NSCI	US President Executive Order establishing the National Strategic Computing Initiative
OPA	Omni Path technology
PCP	Pre-Commercial Procurement
PFlop/s	Peta (= 10^{15}) Floating point operations (usually in 64-bit, i.e. DP) per second, also PF/s.
PRACE	Partnership for Advanced Computing in Europe; Project Acronym
RI	Research Infrastructure
RIA	Research and Innovation Action (type of H2020 project)
R_{max}	Top500 system measured (LINPACK) maximum performance
R_{peak}	Top500 system theoretical maximum performance
SC	Supercomputing Conference; US equivalent to the European based ISC conference. Held annually in U.S.
SKU	Stock Keeping Unit
SoC	System on a Chip
SRA	Strategic Research Agenda
TB	Technical Board (group of Work Package leaders)
TB	Tera (= 2^{40} ~ 10^{12}) Bytes (= 8 bits), also TByte
TCO	Total Cost of Ownership. Includes recurring costs (e.g. personnel, power, cooling, maintenance) in addition to the purchase cost.
TFlop/s	Tera (= 10^{12}) 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

List of Project Partner Acronyms

BADW-LRZ	Leibniz-Rechenzentrum der Bayerischen Akademie der Wissenschaften, Germany (3rd Party to GCS)
BSC	Barcelona Supercomputing Center – Centro Nacional de Supercomputacion, Spain
CaSToRC	Computation-based Science and Technology Research Center, Cyprus
CEA	Commissariat à l'Energie Atomique et aux Energies Alternatives, France (3rd Party to GENCI)
CINECA	CINECA Consorzio Interuniversitario, Italy
CSC	CSC Scientific Computing Ltd., Finland
EPCC	EPCC at The University of Edinburgh, UK
GCS	Gauss Centre for Supercomputing e.V.
GENCI	Grand Equipement National de Calcul Intensiv, France
GRNET	Greek Research and Technology Network, Greece
IT4I	IT4Innovations National Supercomputing Center of VSB-TUO
IUCC	Inter University Computation Centre, Israel
JUELICH	Forschungszentrum Juelich GmbH, Germany
KTH	Royal Institute of Technology, Sweden (3 rd Party to SNIC)
LiU	Linkoping University, Sweden (3 rd Party to SNIC)
NIIF	National Information Infrastructure Development Institute, Hungary
PRACE	Partnership for Advanced Computing in Europe aisbl, Belgium
PSNC	Poznan Supercomputing and Networking Center, Poland
SNIC	Swedish National Infrastructure for Computing (within the Swedish Science Council), Sweden
SURFsara	Dutch national high-performance computing and e-Science support center, part of the SURF cooperative, Netherlands
USTUTT-HLRS	Universitaet Stuttgart – HLRS, Germany (3rd Party to GCS)
VSB-TUO	Vysoka Skola Banska - Technicka Univerzita Ostrava, Czech Republic

Executive Summary

The PRACE-4IP Work Package 5 (WP5), “Best Practices for HPC Systems Commissioning”, has three objectives:

- Procurement independent vendor relations and market watch (Task 1)
- Best practices for energy-efficient HPC Centre Infrastructures design and operations (Task 2)
- Best practices for prototype planning and evaluation (Task 3)

This Work Package builds on the important work performed in previous PRACE Projects in terms of technology watch, know-how and best practices for energy-efficient HPC Centre Infrastructures design and operations, and prototyping of HPC systems. It aims at delivering information and guidance useful for decision makers at different levels. Among them, PRACE aisbl and PRACE sites general managers are clear targets for the technology and market information and orientations collected in this deliverable, but all technical specialists of PRACE partners can be interested in some of the material collected.

This deliverable is the first one of PRACE-4IP Work Package 5 Task 1, it corresponds to a periodic annual update on technology and market trends. This Task 5.1, “Procurement independent vendor relations and market watch”, corresponds to the first objective of Work Package 5. It is thus the continuation of a well-established effort, using assessment of the HPC market based on market surveys, supercomputing conferences, and exchanges with vendors and between experts involved in the work package. Trends and innovations based on the work of prototyping activities in previous PRACE projects are also exploited, as well as the observation of current or new technological R&D projects, such as the PRACE-3IP PCP, the Human Brain Project PCP, FP7 Exascale projects and Horizon 2020 FETHPC1-2014 and follow-ups in future Work Programmes.

1 Introduction

The PRACE-4IP Work Package 5 (WP5), “Best Practices for HPC Systems Commissioning”, has three objectives:

- Procurement independent vendor relations and market watch (Task 1),
- Best practices for energy-efficient HPC Centre Infrastructures design and operations (Task 2),
- Best practices for prototype planning and evaluation (Task 3).

This Work Package builds on the important work performed in all previous PRACE Projects [1] in terms of technology watch, know-how and best practices for energy-efficient HPC Centre Infrastructures design and operations, and prototyping of HPC systems. It aims at delivering information and guidance useful for decision makers at different levels.

Task 5.1 of PRACE-5IP, “Procurement independent vendor relations and market watch”, corresponds to the first objective of Work Package 5. It is the continuation of a well-established effort, using assessment of the HPC market based on market surveys, Top500 analyses, supercomputing conferences, and exchanges with vendors and between experts involved in the work package. Trends and innovations based on the work of prototyping activities in previous PRACE projects are also exploited, as well as the observation of current or new technological R&D projects, such as the PRACE-3IP PCP, the Human Brain Project PCP, FP7 Exascale projects and Horizon 2020 FETHPC1-2014 and follow-ups in future Work Programmes.

This is the first deliverable from Task 1 of Work Package 5 of PRACE. It focussed on technology and market watch only: this means that some best practice and state-of-the-art aspects which were sometimes intertwined with technology watch in past deliverables are now dealt with in other deliverables or white papers (and tasks) of WP5.

This deliverable may contain quite a lot of technical detail on some topics, 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.

This deliverable will be updated by another similar report in one year. It is organised in 5 main chapters. In addition to the introduction (Chapter 1) and the conclusions (Chapter 7) it contains:

- Chapter 2: “Worldwide HPC landscape and market overview” first uses Top500, analysed with a geographical, business topical angle, then proposes some extra considerations from other sources, as well as a brief overview of large HPC initiatives world-wide
- Chapter 3: “Core technologies and components” is a quick overview of processors, accelerators, memory and storage technologies, interconnect technologies
- Chapter 4: “Solution and architectures” gives some vendor snapshots, and looks at some trends in storage, cooling and virtualisation and cloud delivery.
- Chapter 5: “Management tools” is an overview of various tools at system, user resources or log analysis levels.
- Chapter 6: “EU Projects for Exascale and Big Data” scans FP7, H2020 and other projects towards exascale.

2 Worldwide HPC landscape and market overview

2.1 A quick snapshot of HPC worldwide

The purpose of this section is to present an overview of HPC worldwide, with a special focus on Europe, based on statics derived from the Top500, the Green 500 and the HPCG benchmarks [2][3][4][5].

The focus here is on large systems. Therefore, each time such analysis is applicable and meaningful, special attention is paid to systems among the 50 most powerful in the world according to the Top500 ranking (called "Top50" hereafter). This choice was made according to the fact that, in most cases:

- a Tier-0 system is part of the 50 most powerful system in the world during most of its lifetime,
- a Tier-1 system is part of Top 50 at the time of its installation.

2.1.1 *Countries*

Analysis was done for the countries that were at least once part of the 10 largest countries in terms of cumulated R_{max} in the Top50 in the past 5 years:

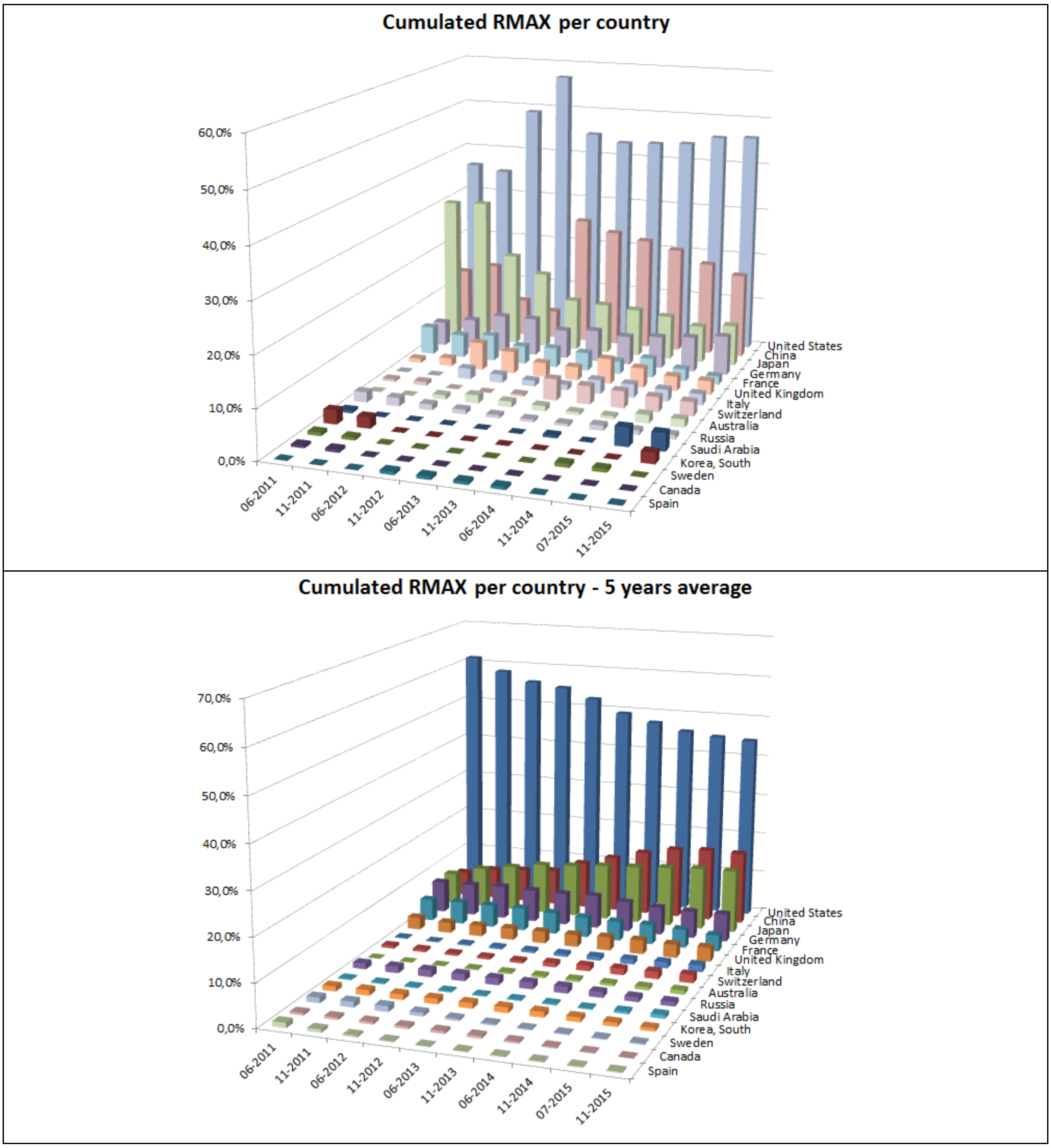
Country	Cumulated R_{max} in Top50 (06-2011 to 11-2015)		
	Number of times part of the 10 largest countries	Best ranking reached	Number of times in the best position
United States	10	1	10
China	10	2	6
Japan	10	2	4
Germany	10	3	2
France	10	4	1
United Kingdom	10	5	4
Italy	8	7	4
Switzerland	6	5	1
Australia	6	8	2
Russia	5	7	2
Saudi Arabia	3	5	2
Korea, South	3	6	2
Sweden	2	9	1
Canada	1	10	1
Spain	1	10	1

Table 1: countries for Top50 analysis

Figure 1 and Figure 2 below show the evolution of the fraction of cumulated Rmax overtime for these countries derived from the Top500. Two figures are shown: one being the fraction of cumulated Rmax for a specific Top500 edition, one being the average over the 5 previous years for a specific Top500 edition. The 5-year period allows to smooth the fluctuations; it was chosen since it is the typical lifetime of a supercomputer.

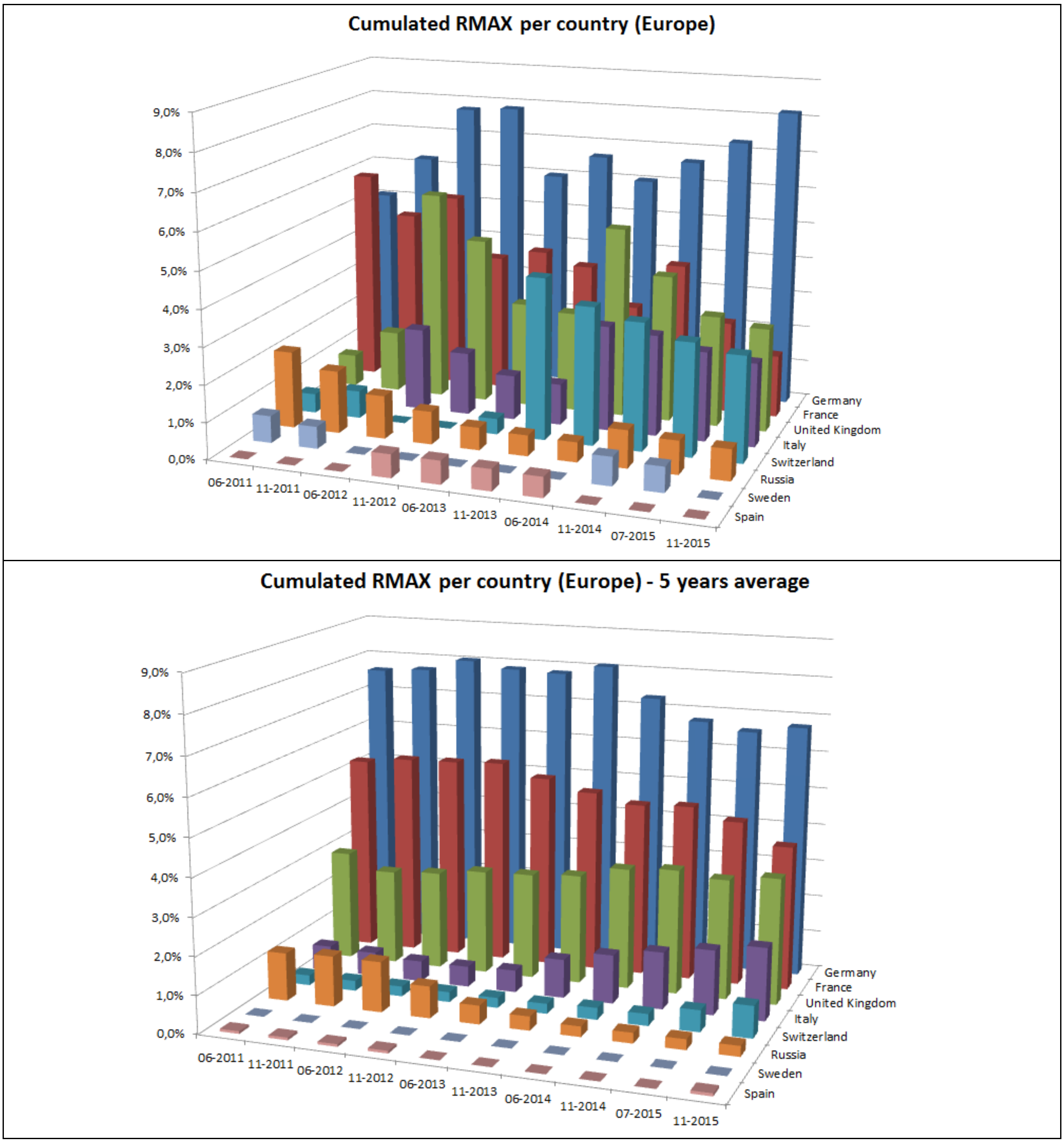
Regarding the world, the figure shows that, for the previous 18 months, the position of United States and China tends to stabilize after a period during which there was a strong decrease of the United States and a strong increase of China. The other big players are Japan and several European countries.

Regarding Europe, Germany is heading the race while France, United Kingdom, Italy and Switzerland are close to each other with a decreasing slope for France, increasing for the others.


Cumulated RMAX per country

Cumulated RMAX per country - 5 years average

Figure 1: evolution of the fraction of cumulated RMAX overtime for top HPC countries



Cumulated RMAX per country (Europe)

Cumulated RMAX per country (Europe) - 5 years average

Figure 2: evolution of the fraction of cumulated RMAX overtime for top HPC countries / Europe

Figure 3 below focuses on Europe in terms of number of systems in the Top10/20/50. The figure shows clearly that the largest systems are not in Europe.

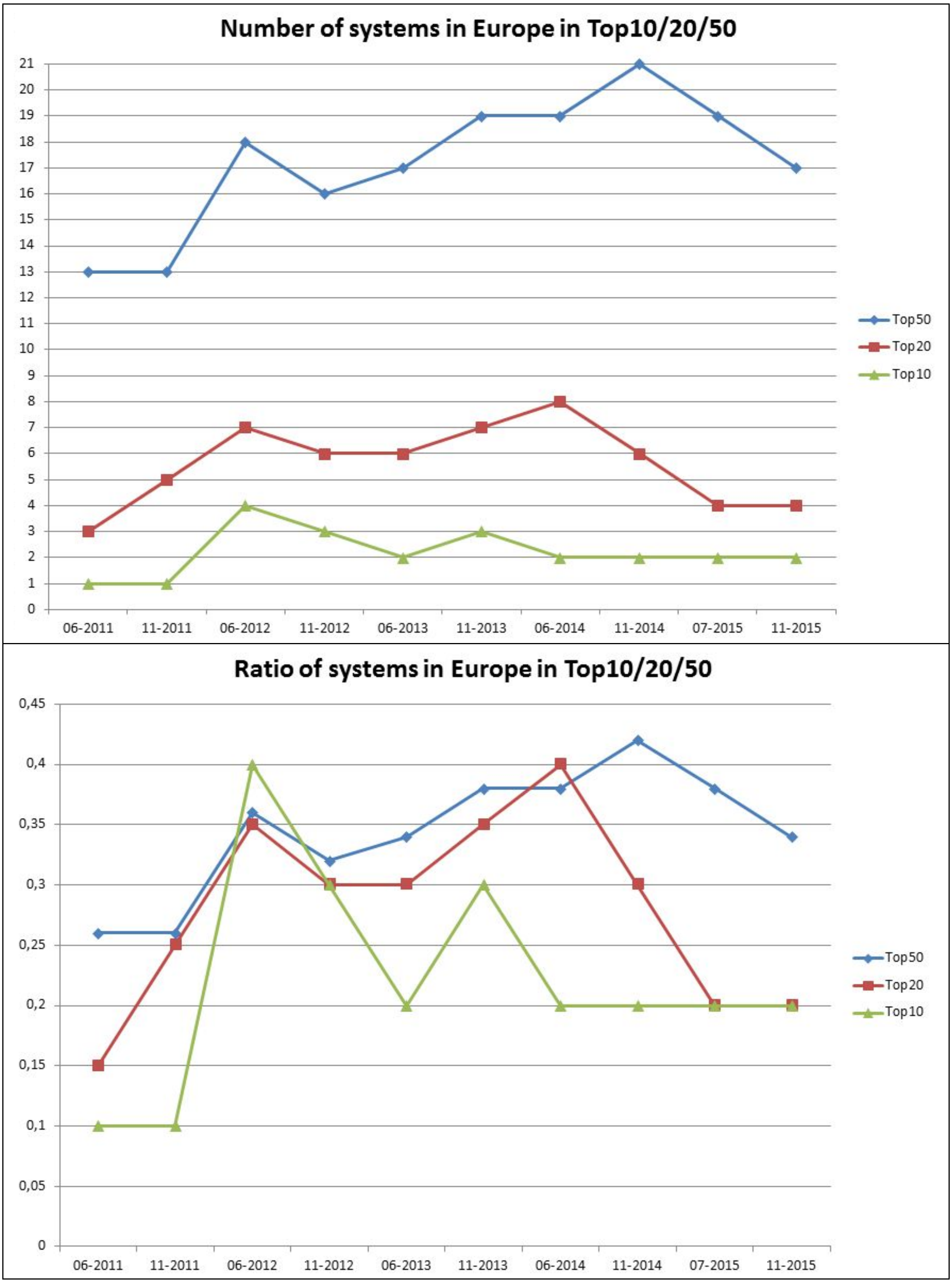


Figure 3: Top 10/20/50 systems in Europe

2.1.2 *Accelerators*

Figure 4 below, on the left, shows the fraction of systems equipped with accelerators in the Top50. The figure on the right compares Europe with the world in term of accelerators both for the Top50 and the Top500.

Regarding the first figure, the growth of fraction of system equipped with accelerators tends to slow down. Nvidia GPU and Intel Xeon Phi are the major players and are quite comparable in terms of penetration, with a slight advantage for Nvidia GPU.

Regarding the second figure, the fraction of accelerators is roughly the same in the Top50 and in the Top500. The penetration of Intel Xeon Phi in Europe is much lower than in the world, Nvidia GPU being by far in Europe (both in Top50 and Top500) the leader.

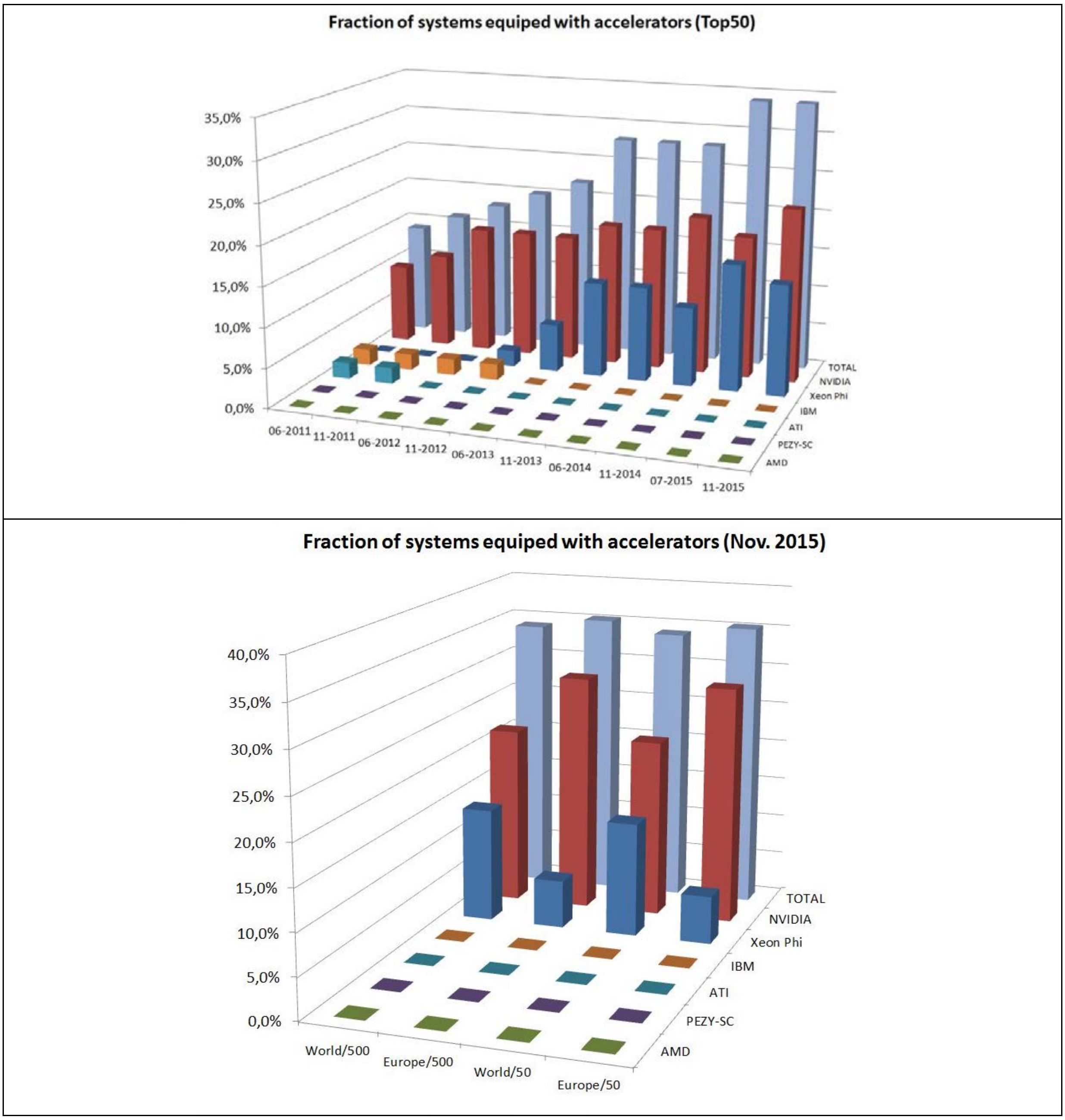


Figure 4: Fraction of systems with accelerators

2.1.3 *Age*

Figure 5 below shows the age of systems (in terms of time of presence in the Top500) for the world and for Europe.

The age of systems has been steadily increasing for the last 5 years for the Top50 and the Top500 except for the last year during which the average age of system in the Top50 has slightly decreased at the world level.

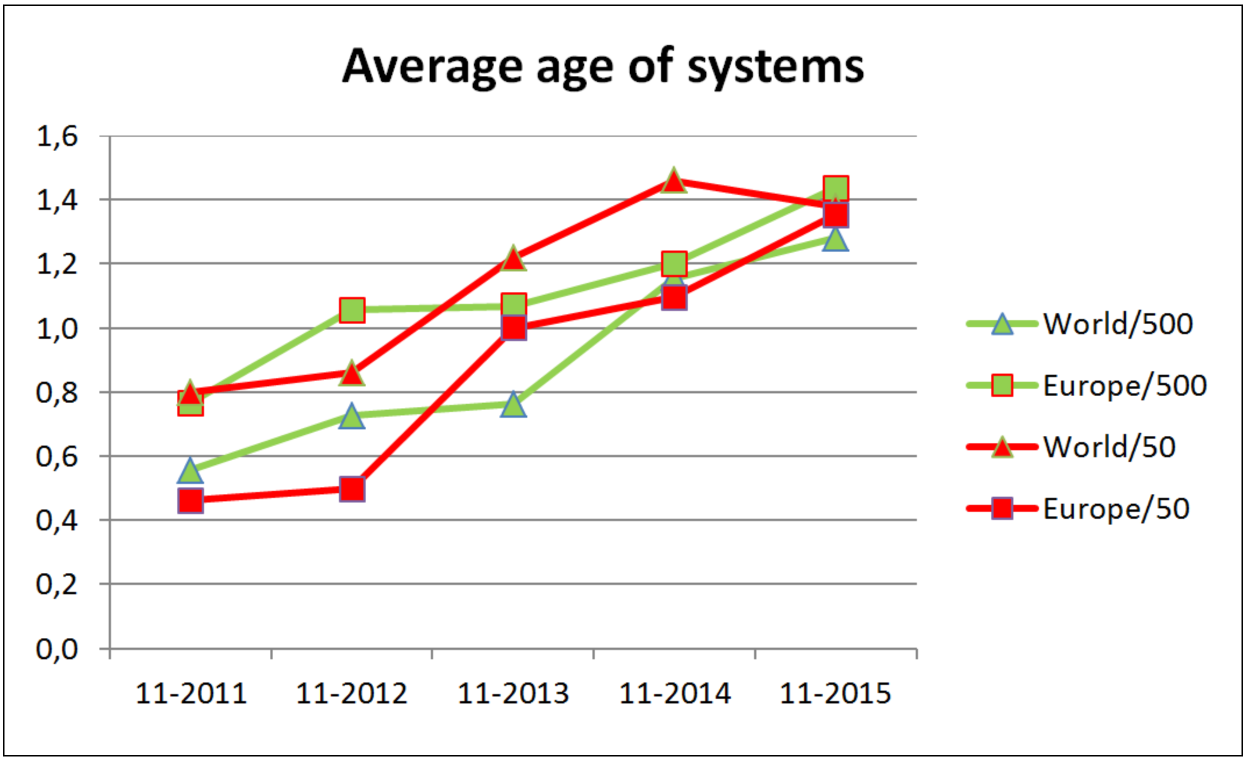
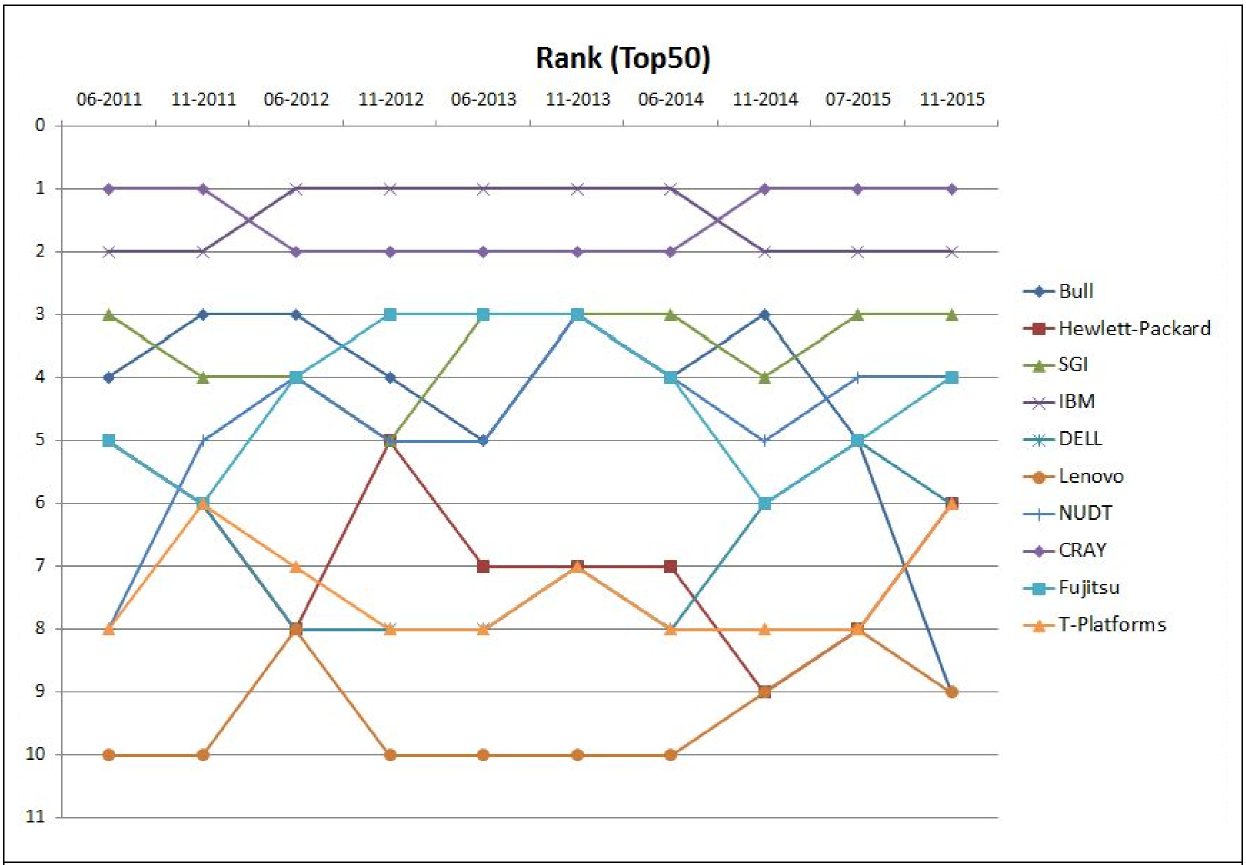
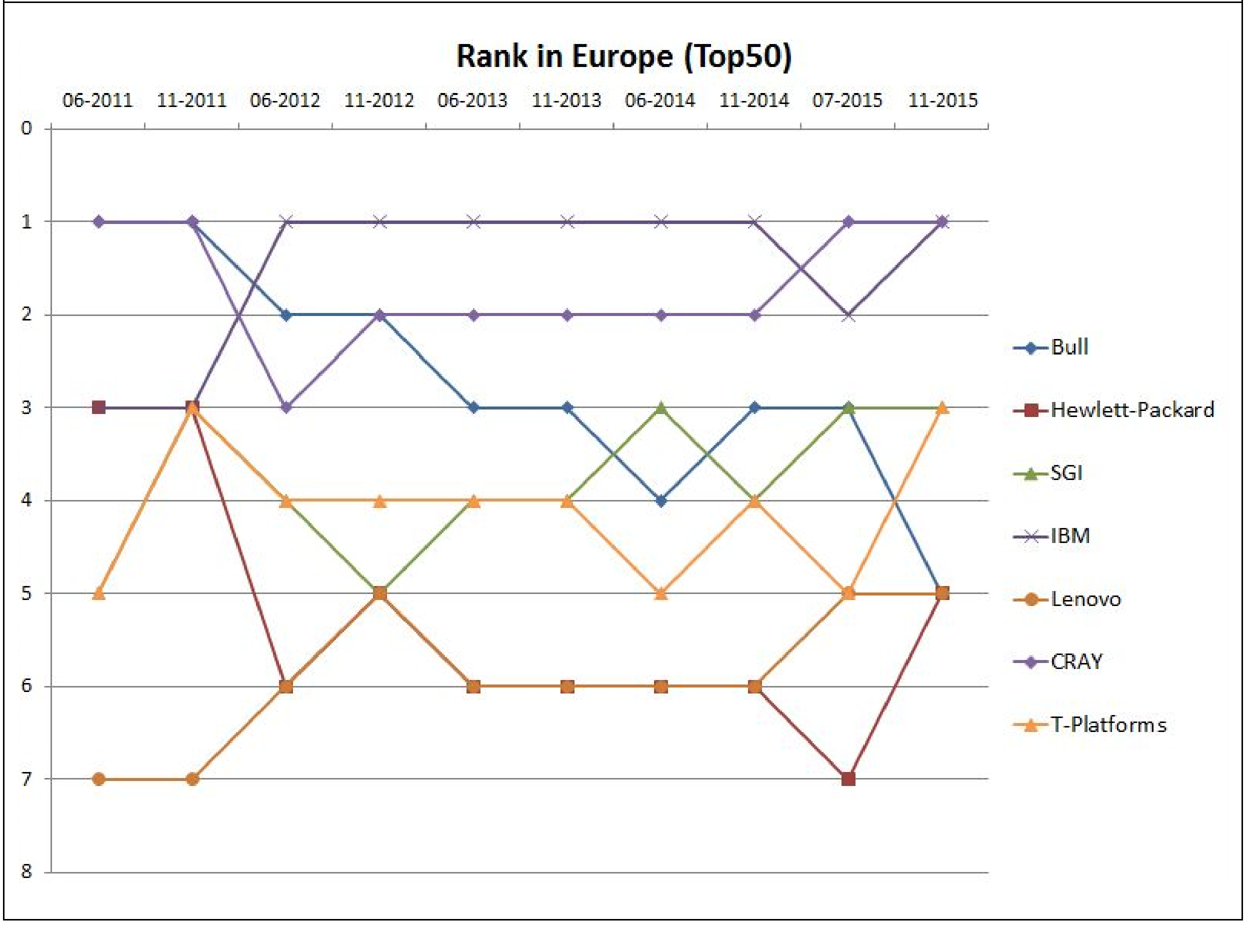


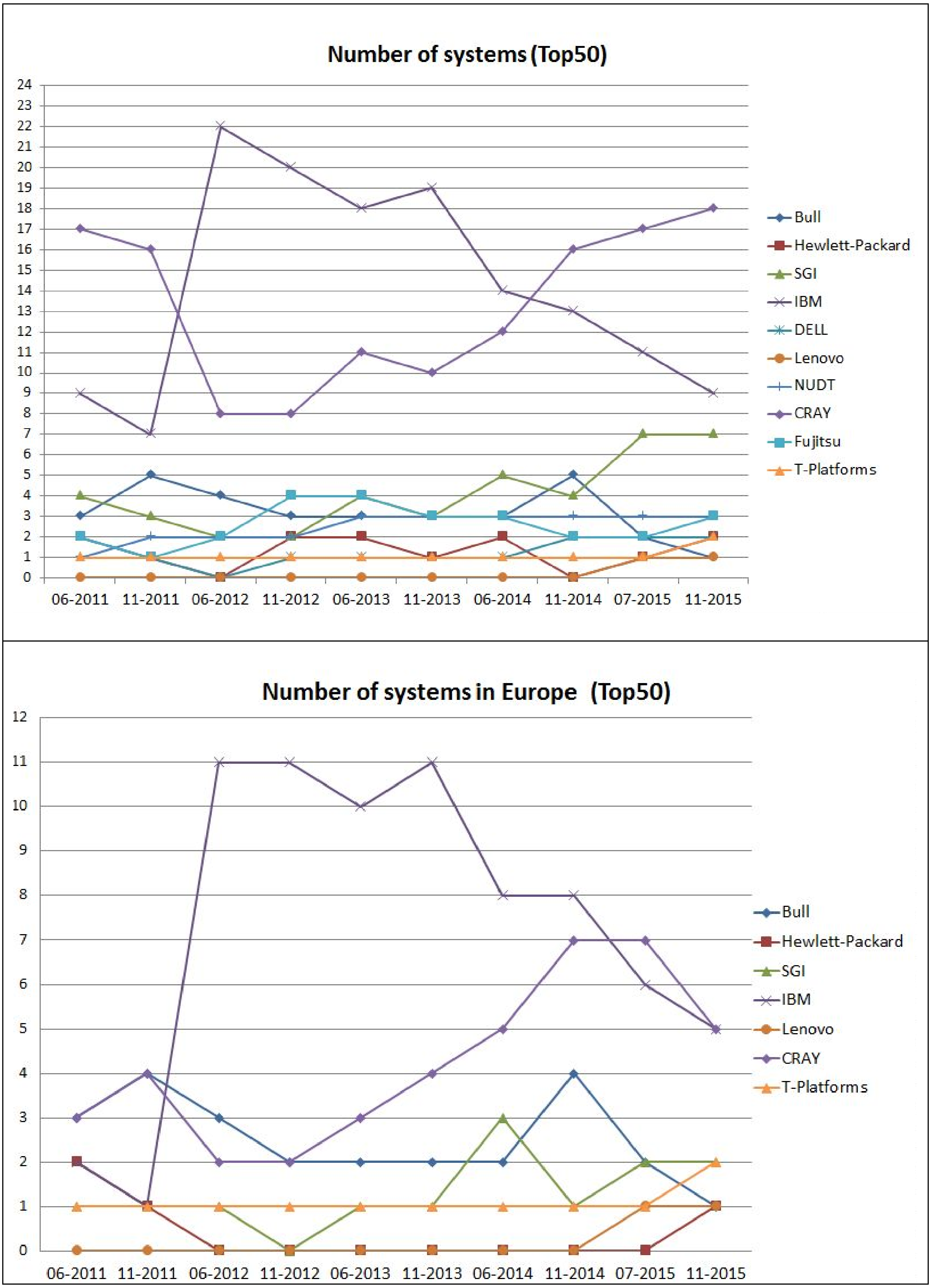
Figure 5: Average age of systems

2.1.4 *Vendors*

The charts in Figure 6 below show the relative position of the vendor at the world and European level. The most visible trend is the fall of IBM, related to the fact that IBM is no longer selling x86 HPC systems (LENOVO taking over for this activity), and the rise of Cray. This trend is visible at European level and stronger at the world level.


Rank (Top50)
06-2011
11-2011
06-2012
11-2012
06-2013
11-2013
06-2014
11-2014
07-2015
11-2015
0
1
2
3
4
5
6
7
8
9
10
11
Bull
Hewlett-Packard
SGI
IBM
DELL
Lenovo
NUDT
CRAY
Fujitsu
T-Platforms


Rank in Europe (Top50)
06-2011
11-2011
06-2012
11-2012
06-2013
11-2013
06-2014
11-2014
07-2015
11-2015
0
1
2
3
4
5
6
7
8
Bull
Hewlett-Packard
SGI
IBM
Lenovo
CRAY
T-Platforms



Number of systems (Top50)

Number of systems in Europe (Top50)

Figure 6: Top500 vendors

Figure 7 below focus on the only European vendor at Top50/Tier-0 capability level (excluding vendors from Russia): Bull. They show that the presence of Bull in the Top50 has recently decreased, the presence is stable in the Top100 and increasing in the Top500.

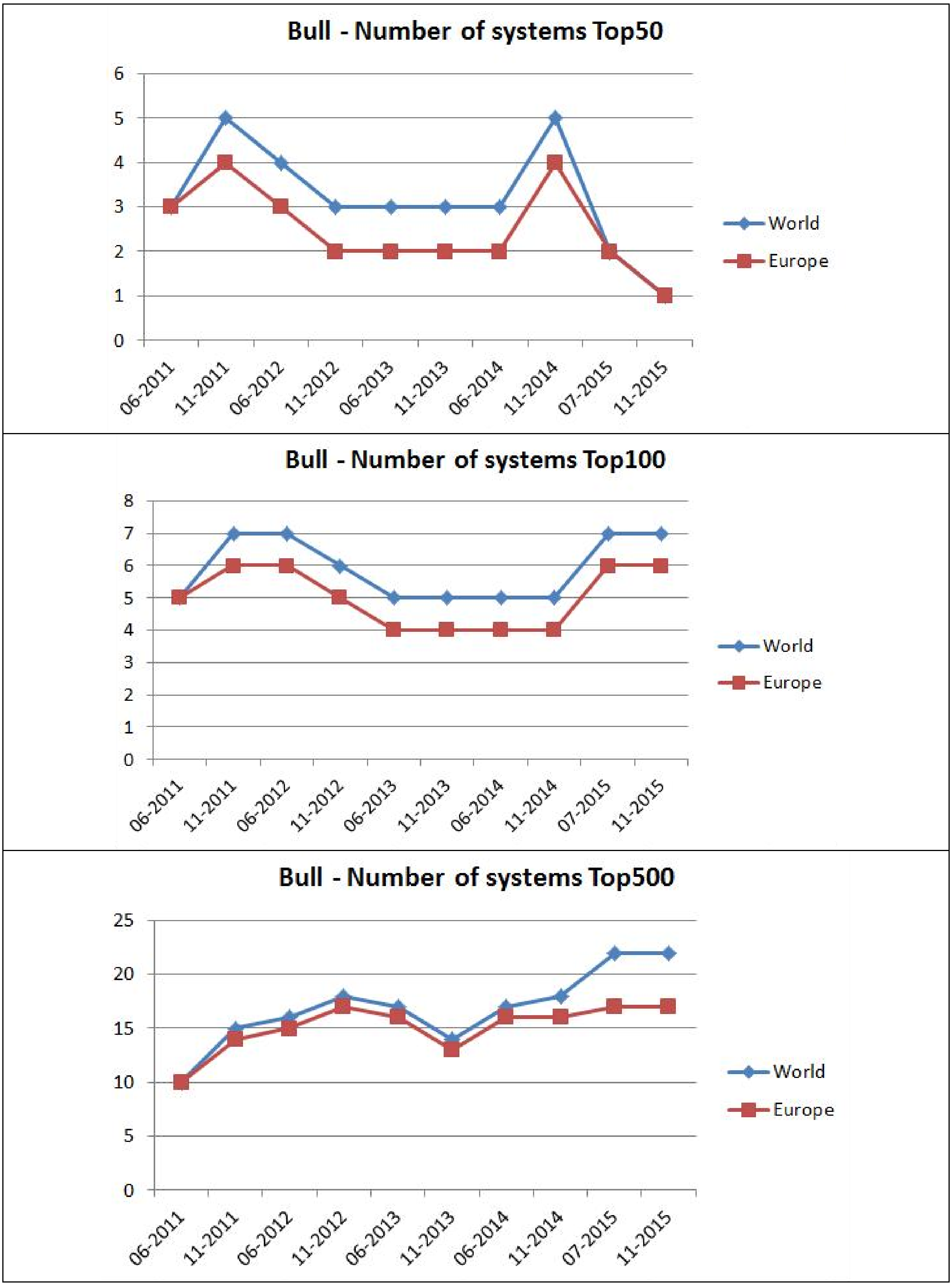
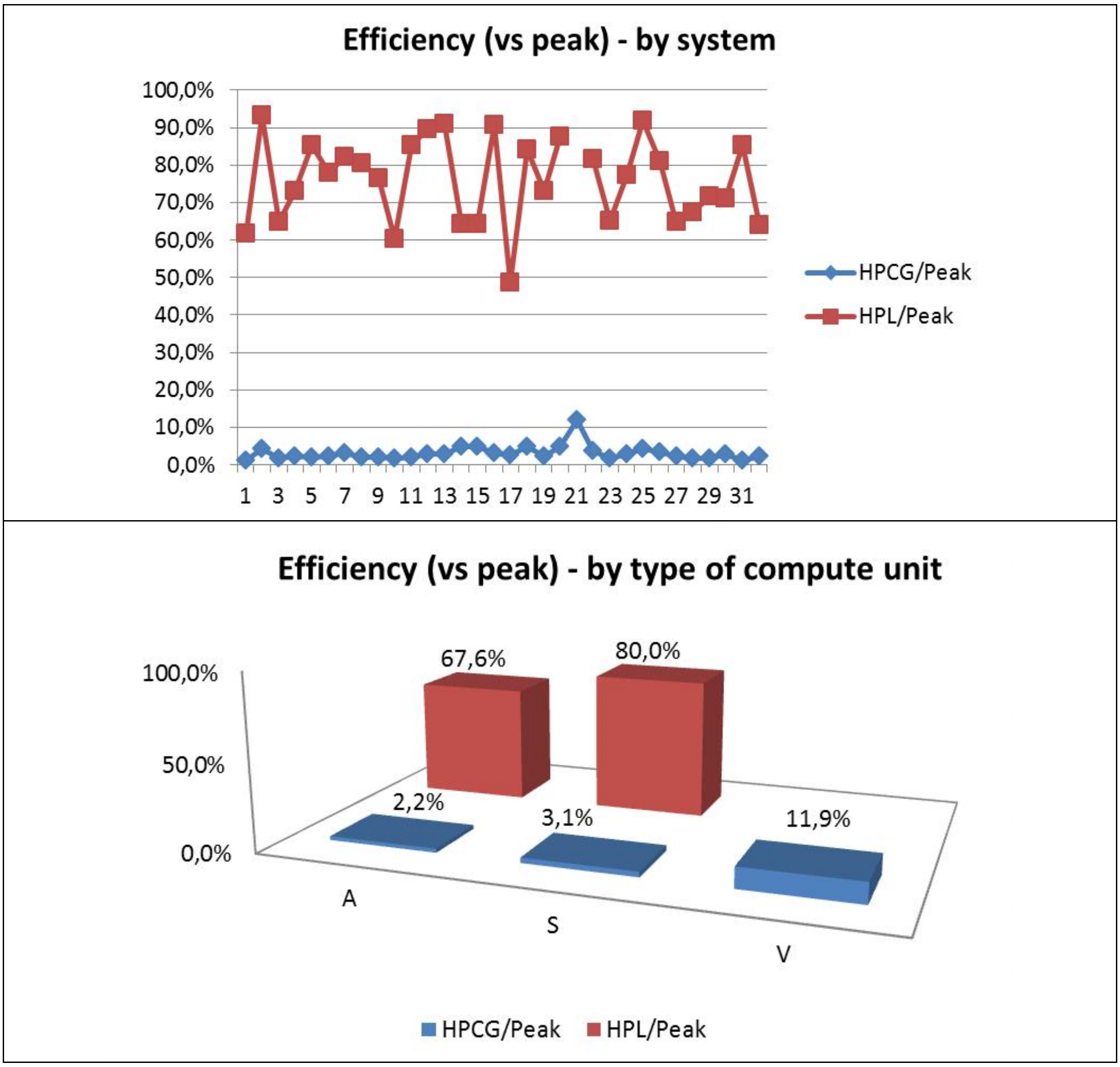


Figure 7: Top500 European vendor

2.1.5 *Computing efficiency*

In Figure 8 HPCG shows a much smaller variation than HPL, with the exception of system #21. This system is a NEC SX-ACE based system (Earth Simulator at CEIST), and the same pattern applies to all the SX-ACE systems on the HPCG list. This architecture seems to score well on the HPCG benchmark relative to HPL.



On the top figure the systems are sorted by HPCG rank (from #1 to #32)
On the bottom figure, A= accelerated, S= scalar, V=vector.

Figure 8: HPL vs. HPCG efficiency comparison

2.1.6 *Energy efficiency*

While the Green500 efficiency ratio is steadily climbing upwards, the Top10 systems are showing a slight decrease. As seen in the second graph of Figure 9 this is not the case when widening the base to Top50. This reflects a more marked focus on peak performance for the high end and a widening gap between R_{max} and R_{peak}. Between 2013 and 2015 several energy efficient systems exited the Top10.

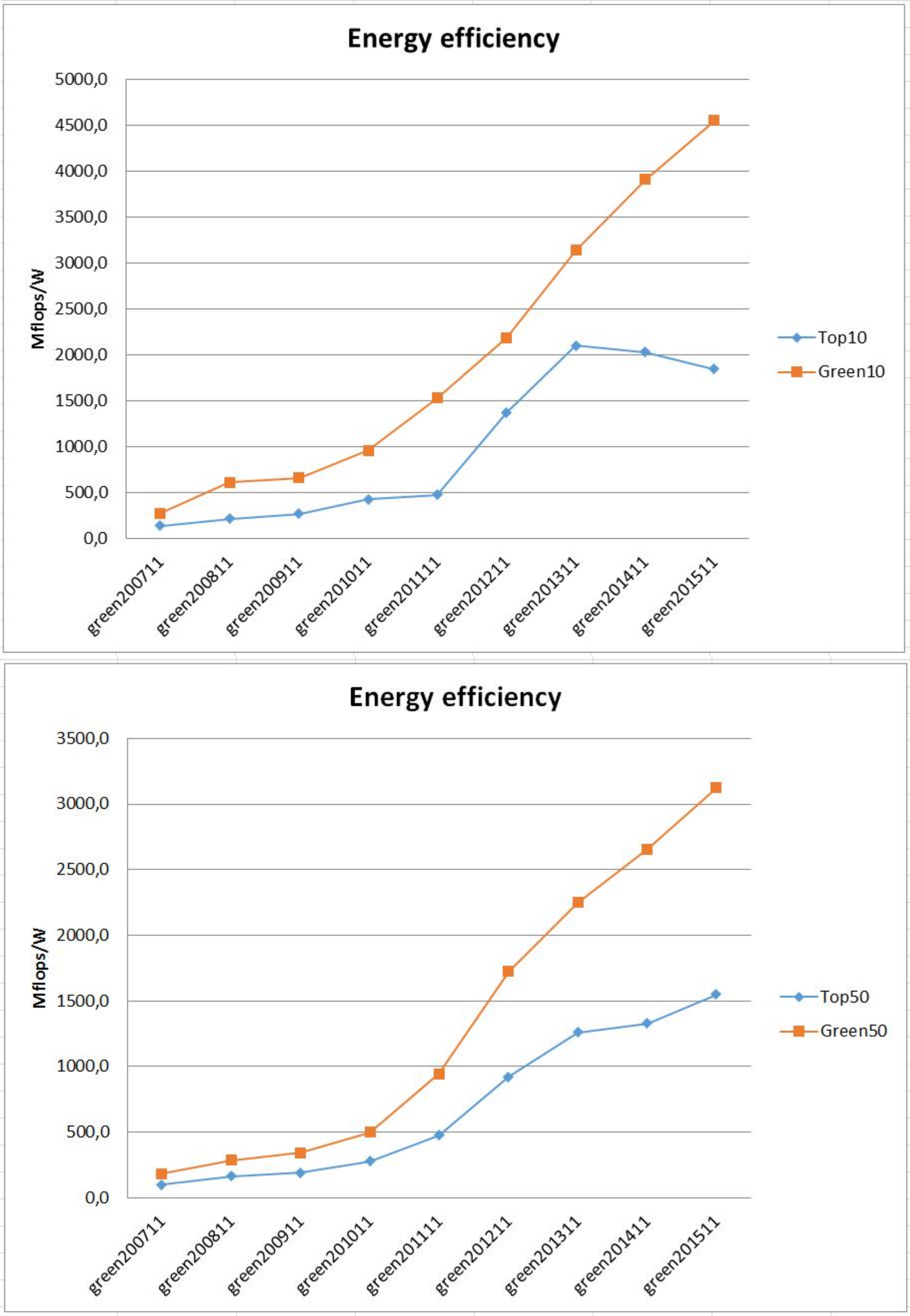


Figure 9: Energy efficiency in Top10/50 (Top500 and Green500 references)

2.2 Some more business analysis

2.2.1 *General impressions from SC'15*

Intel was dominating the conference and had a very high visibility, also through collaborations with other vendors. Some vendors even seemed to keep their non-Intel based products out of the limelight due to this. Not much in the way of products was shown, that wasn't based on Intel hardware on the CPU level.

Storage played a more important role than ever, along with data analysis. This is of course fuelled by the "Big Data" trend.

Chinese vendor Sugon (formerly Dawning) made a large showing on the TOP500, but this was mainly due to putting in the effort to benchmark already installed systems. There have been some press releases about Sugon and its primary Chinese competitor Inspur, both having a strong presence at SC15, about their possible new US and Europe go-to-market strategy [6].

2.2.2 *Market trends according to some analysts*

There is a converging vision and analysis of the market analysts on a few facts and projections [7][8][9][10][11]:

- good overall shape of the HPC and servers market, especially in Asia/Pacific and with a strong China lead that is pulling the growth (top four Chinese OEMs – Inspur, Huawei, Lenovo, and Sugon); high-end of the sever market, which is roughly the HPC segment, has less robust and more irregular growth
- HPE a quantitative leader, all the more than the usual challenger, IBM, which has now split – LENOVO and IBM shares in total roughly accounting for past IBM alone share
- still dominant x86 position, IBM POWER server sales growing, ARM server sales not growing
- storage, HPDA (High Performance Data Analytics), rising in terms of market volume

Some of these trends merely confirm the impressions and observations collected during SC15, as explained in previous section.

2.3 Large initiatives world-wide

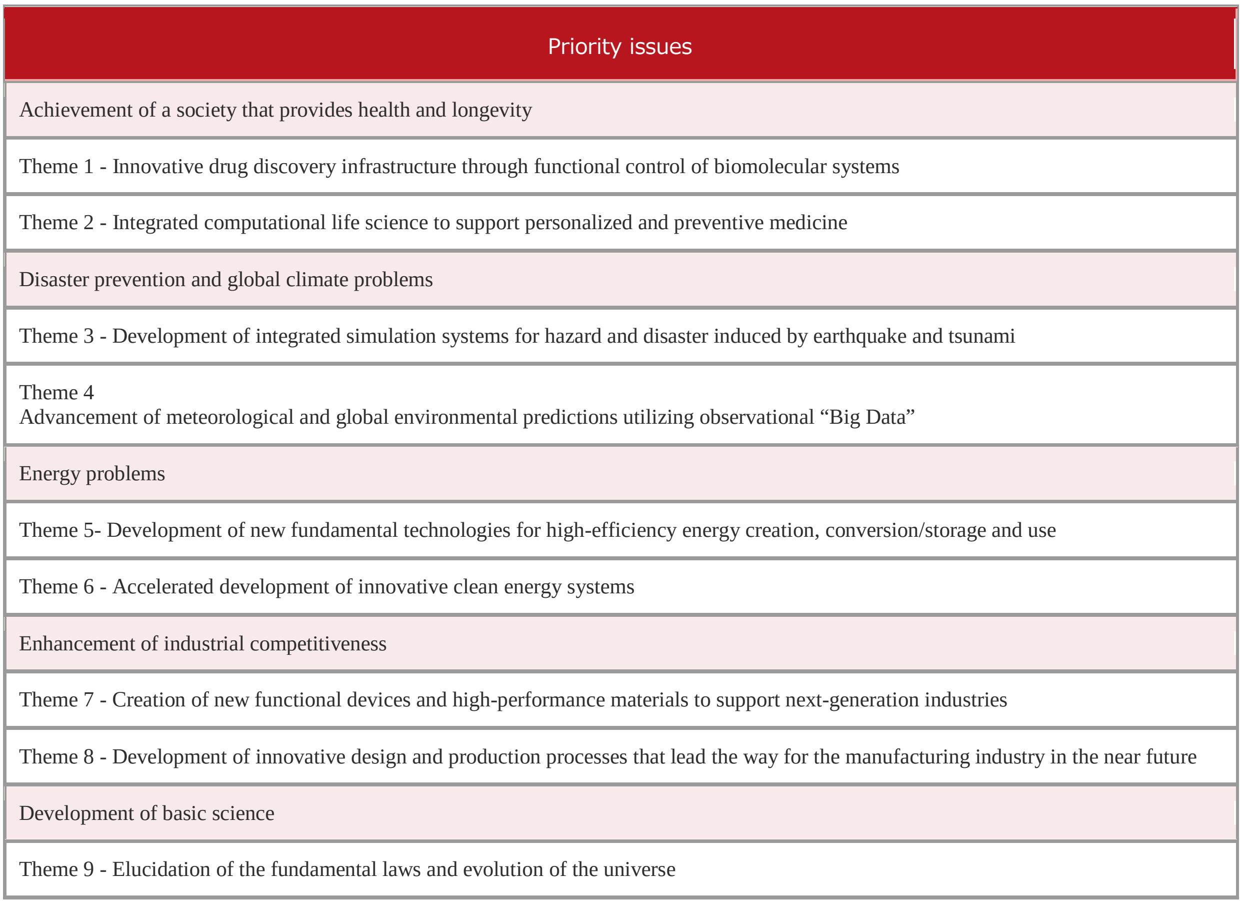
2.3.1 *US NSCI*

In July 2015 President Obama issued an Executive Order establishing the National Strategic Computing Initiative (NSCI) aiming at ensuring the United States continues leading in this field over the coming decades [12][13]. NSCI defines a multi-agency framework for furthering U.S. "economic competitiveness and scientific discovery" through orchestrated advances in high performance computing (HPC). We can see this as a further step in US efforts to structure and fund a federal initiative in the area of HPC – in the last decades we had observed a number of different programmes from DARPA or DOE, with some progressive connection and alignment of the policies and approaches of different labs or agencies (the DOE/NNSA with a three-lab approach, then DOE NNSA and Office of Science together - CORAL joint Collaboration of Oak Ridge, Argonne, and Lawrence Livermore (CORAL) was established in early 2014). CORAL procurement is typical of US approach, with a single process to acquire next-generation supercomputers with strong R&I requirements for three DOE national laboratories [15]. As IDC states [18][33]: "Government and government-related portions of the U.S. HPC market are effectively closed to non-U.S. system vendors by "Buy American" legislation and other preferential policies and practices".

This time the global economic impact of the NSCI programme, boosting innovation and expanding HPC usage to industrial and societal usages, are particularly emphasized. The announcement has been widely commented since mid-2015 and a recurring discussion has been on funding amounts and sustainability – the US political mechanisms and agenda make the funding process complex and progressive. This goes with a revised target of exascale machines by 2023.

2.3.2 *Japan Post K and Flagship 2020*

Japan launched in 2014 a new flagship HPC project for the 2020 horizon - now called the FLAGSHIP 2020 Project - targeting exascale for the succession of K computer in 2020. MEXT - Ministry of Education, Culture, Sports, Science and Technology - is still the driving authority, with RIKEN AICS (Advanced Institute for Computational Science) in charge of the implementation of the dedicated project, which encompasses technology development and related application enabling. There was a tender leading to the selection of Fujitsu as the vendor in charge [19][20][21][22]. Kobe AICS site would host the future system. Budget is in the order of 850 M€ (110 billion yens), including R&D, machine acquisition and application development. Co-design is strongly asserted between RIKEN and Fujitsu. Scientific and societal challenges have already been carefully selected with organisations in charge, covering 9 themes in 5 areas as shown in table Table 2 below.



Priority issues
Achievement of a society that provides health and longevity
Theme 1 - Innovative drug discovery infrastructure through functional control of biomolecular systems
Theme 2 - Integrated computational life science to support personalized and preventive medicine
Disaster prevention and global climate problems
Theme 3 - Development of integrated simulation systems for hazard and disaster induced by earthquake and tsunami
Theme 4 Advancement of meteorological and global environmental predictions utilizing observational “Big Data”
Energy problems
Theme 5- Development of new fundamental technologies for high-efficiency energy creation, conversion/storage and use
Theme 6 - Accelerated development of innovative clean energy systems
Enhancement of industrial competitiveness
Theme 7 - Creation of new functional devices and high-performance materials to support next-generation industries
Theme 8 - Development of innovative design and production processes that lead the way for the manufacturing industry in the near future
Development of basic science
Theme 9 - Elucidation of the fundamental laws and evolution of the universe

Table 2: Social and scientific priority issues to be tackled by using Post K computer

The term FLAGSHIP is also an abbreviation of Future LAtency core-based General-purpose Supercomputer with HIgh Productivity, which shows the strong concern for usability and targeted pervasiness of usage of outcomes of the project.

2.3.3 *China*

China has had a steady HPC ambition, primarily via government entities Chinese Academy of Sciences (CAS) and its academic partner the National University of Defense Technology (NUDT) for quite a while, even including the development of domestic processors. Recent economic slowing down in China may affect this but probably not change the mid to long-term trend. Chinese government investments in HPC have risen rapidly in recent years, and the willingness of independence from non-Chinese (primarily U.S.)

Last year there were ‘diplomatic’ moves between the US and China, U.S. government blocking Intel from exporting its processors to upgrade Tianhe-2 and some other leading supercomputers in China - claiming that the computers were being used for nuclear weapons research. This probably exacerbated or at least did not slow down Chinese plans for developing indigenous technology [16].

The U.S. government's recent decision to ban Intel from exporting Xeon x86 and Xeon Phi processors to upgrade Tianhe-2 and several other leading supercomputers in China may accelerate China's efforts (and financing) to advance domestic processor development. Also of note, China has urged investment banks and other "critical infrastructure" sites to replace non-Chinese HPC systems (primarily IBM) with Chinese supercomputers. Lenovo's acquisition of IBM's x86 server business should enable Lenovo as a Chinese firm to move into these sites. Some Chinese supercomputers already incorporate Chinese processors. Work is on-going on the follow-up of Tianhe-2 Express-2 interconnect network, compliant with Intel Phi but also a domestic DSP chip now announced by the National University of Defence – the Matrix 2000 GPDSP ‘China Accelerator’, a 64-bit, 2.4 Teraflops one (in double precision) running at 1 GHz with a power consumption of around 200W –comparable to the Intel Xeon Phi [17].

2.3.4 *Europe*

Horizon 2020 HPC programme is now well established, with the contractual Public Private Partnership (cPPP) encompassing technology and related R&D, and Centres of Excellence for Computing Applications (CoEs) [23]. Nineteen so-called FETHPC projects were selected in 2015, plus two Coordination and Support Actions (CSA) [131]. Besides this eight Centres of Excellences [119] were also selected from the calls launched in the framework of Work Programme 2014-2015.

Chapter 6 gives more details on these projects, with some quick technical hints.

The ‘HPC infrastructure’ pillar is not formally included in the scope of the cPPP (which covers Technologies – FETHPC – and Applications – CoEs), but has strong links developed at PRACE aisbl and PRACE IP projects levels. PRACE aisbl [1] is still working on the transition to ‘PRACE 2” after the successful initial period of 5 years of the infrastructure. PRACE 3IP project is still running via its PCP activity – see section 6.6.1- in parallel of PRACE 4IP. A PRACE 5IP project has just been submitted in EINFRA-11-2016 call closing end of March [24].

FETHPC and CàE calls within Work Programme 2014-2015 of H2020 totalised around 140 M€ of EC funding, out of the 700 provisioned for the HPC cPPP.

ETP4HPC [28] has been growing steadily (gaining more members, 72 as of March 2016 [29]) and produced a new release of its Strategic Research Agenda (SRA) in November 2016 [27]. This document is the more detailed technical reference going with the HPC call in Work

Programme 2016-2107, which was elaborated using the EC/ ETP4HPC structured dialogue within the cPPP framework.

The FETHPC call in Work Programme 2016-2017has 3 topics [25][26]:

- FETHPC-01-2016: Co-design of HPC systems and applications, for RIA (Research and Innovation Action) projects, closing September 2016 with a budget of €41 M€.
- FETHPC-02-2017: Transition to Exascale Computing, for RIA projects, closing September 2017, with a budget of €40 M€.
- FETHPC-03-2017: Exascale HPC ecosystem development, for CSA (Coordination and Support Action) projects, closing September 2017 with a budget of €4 M€.

It can thus be said that the programme is on track and well engaged now. It already shows increased interaction between academia, research organisations, and private companies on the technology supply side (including many SMEs). However, the fragmentation into many relatively small projects remains quite high, even if WP 2016-2107 is showing some focus effort (FETHPC-01-2016 topic).

It must be noted that a significant amount of funding remains to be allocated out of the cPPP 700 M€. Future Work Programme 2018-2020 should thus have higher funding that the two previous ones. ETP4HPC is working on the R&D contents for this future stage, also proposing a concept of Extreme Scale Demonstrators that would integrate outcomes of R&D projects funded in Work Programmes 2014-2015 and 2016-2107 (see ETP4HPC SRA 2015 - chapter 8 in [27]).

FETHPC-1 call of WP206-2017 also funded two CSAs (cf section 6.4). EXDCI, the European Extreme Data & Computing Initiative, is one of these [30]. Its objective is to coordinate the development and implementation of a common strategy for the European HPC Ecosystem. The two most significant HPC bodies in Europe, PRACE and ETP4HPC, joined their expertise in this 30-month project. EXDCI’s first period (first 6 months) activities show very promising momentum for the whole EU HPC community to develop joint visions (Rome Sept 2015 [32], Prague May 2016 [31]).

Mid-2015 IDC report (SMART action funded by the EC [18][33]) was published on the progress of the European HPC Strategy toward ensuring European leadership in the supply and use of HPC systems and services by 2020, and provided recommendations for its implementation. Good progress has been reported regarding the overall progress and momentum, while emphasizing some questions or shortcomings.

A general trend and a framework for analysis is the intertwining of HPC with a growing number of industrial applications and scientific and societal domains. This places HPC as one of the key contributors to the Digital Single Market (DSM) strategy being developed by the EC since more than one year [34]. More precise plans are expected to be made public soon (April) in this perspective, updating and widening 2012 EC Communication "High Performance Computing (HPC): Europe's place in a global race", including the notion of a European Science Cloud [35].

Meanwhile (November 2015) an initiative led by a few Member States (Luxembourg, France, Italy, Spain) has been announced, so-called IPCEI HPC-BDA, an Important Project of Common European Interest mixing HPC and Big Data objectives [36]. This is understood as an action from volunteering countries wanting to optimise different aspects of a large European initiative aligned with the objectives of the aforementioned plan to come. Aggregation of fundings from EC sources, national sources and private sources to support the project, state aids regulations, competition regulation are aspects that such an IPECI can help optimise. Since this is all work in progress and more precise communications to come, this will be further documented and commented in future reports.

In conclusion of this section, it can be said that Europe is clearly willing to move forward in the area of HPC in a wider context of Digitalised Economy. Indeed, if it is often said that Europe is lagging behind, w.r.t. other continents or regions, in terms of HPC budget,this is only partly and very relatively true – the 700 M€ of the HPC cPPP for instance are not a small budget, although quite fragmented in many different projects. So it seems extra budget and funding efforts are now being considered, but, equally important, global strategy and more efficient organisation and structuration.

3 Core technologies and components

3.1 Processors

In the last decade, the trend of commodity-based multicore server processors has become the common architectural choice for HPC systems. While clock speeds have largely stagnated, performance is gained from increasingly large multicore architectures, larger caches and increasingly wider vectorization. It is to be expected that this trend will continue.

The viable set of processor vendors have also shrunk. In this section we discuss the viable processor technologies now and in the near future.

3.1.1 *Intel x86*

In the last 5 years Intel has become clearly the most dominant processor vendor in the HPC market.

Skylake

The next generation microarchitecture of Intel processors is codenamed Skylake and it is expected in mid-2017. The chips will use the same 14nm technology as the Broadwell CPUs. Furthermore their key features include:

- AVX-512 instruction set. Most notably the new set features 32 of 512-bit vector registers (up from 16 and 256bits of AVX-2). This increases the theoretical peak DP FLOP/s rate to 16 per clock cycle. Other new features include new vector mask registers, improved gather/scatter support and conflict detection instructions for loops. Furthermore, the instruction set (with slight differences) is also employed in the Knights Landing processors (see 3.2.1) with binary compatibility promised between Skylake and Knights Landing.
- 6 memory channels (up from 4 in the current Haswell generation). This should help enable to keep the memory bandwidth somewhat balanced with the constantly increasing core count. However, this is dependent ultimately on the core counts of the upcoming processors, details of which have not been officially released at the time of writing.
- New generation NVRAM DIMMs. Two memory channels can be fitted with NVRAM DIMMs which should provide fairly good performance and have much larger capacity than the conventional RAM.
- The Omni-Path 1.0 interconnect will be integrated to the CPU package at least in some SKUs (Stock Keeping Units). This should reduce server footprint, interconnect power consumption and increase the server reliability.

Cannonlake

It is expected that Skylake will be succeeded with a “shrink” based on the 10nm process. This CPU will share the microarchitecture of Skylake but will likely feature more cores.

Beyond Cannonlake

The next microarchitecture update is expected in 2018 and 2019 but little has been officially announced. It should be expected that some of the innovations introduced in the Xeon Phi product line may be introduced to the CPU products.

3.1.2 *AMD x86*

While AMD has been largely marginalized in the HPC processor community and has focused its effort to the consumer market, the processor manufacturer is planning to return to the high-performance server market with the Zen CPU architecture, planned for 2016. The processor will be using the Intel-style Simultaneous Multithreading Architecture (SMT) instead of the Clustered Multithreading approach first introduced in the AMD “Bulldozer” Opteron cores (Figure 1). The processor will be manufactured with the 14nm process. It is still under speculation which exact level of micro-architecture support of the processors it will feature (AVX2 or AVX-512).

Another potentially interesting variant is an “APU” variant with a sizeable integrated GPU complementing the CPU cores.

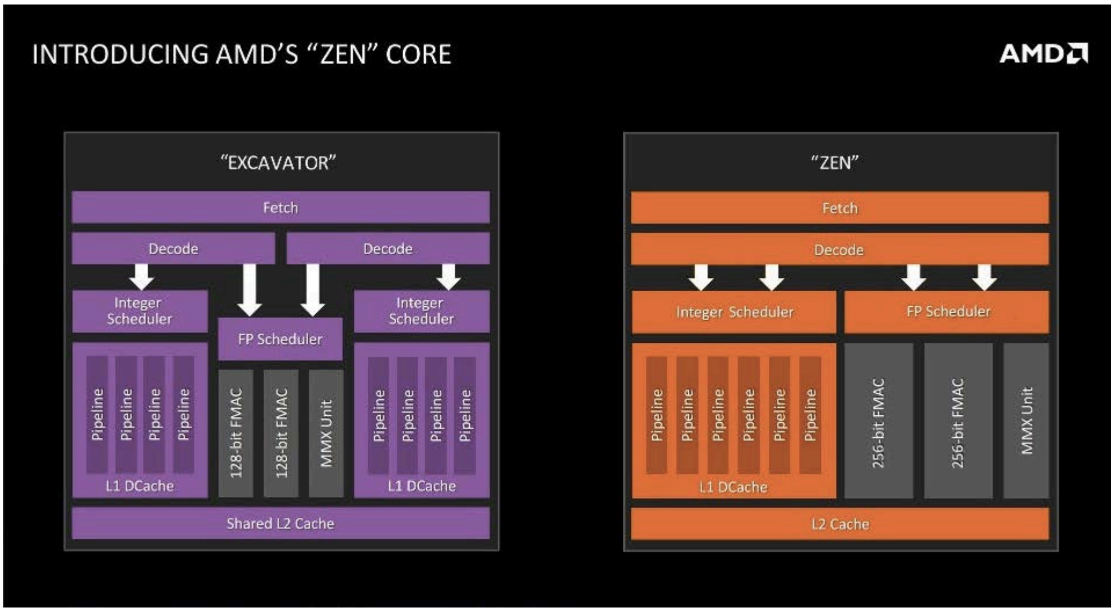

INTRODUCING AMD'S “ZEN” CORE

Figure 10: AMD Zen compared with the previous generation Excavator architecture

3.1.3 *ARM (ecosystem with chip makers)*

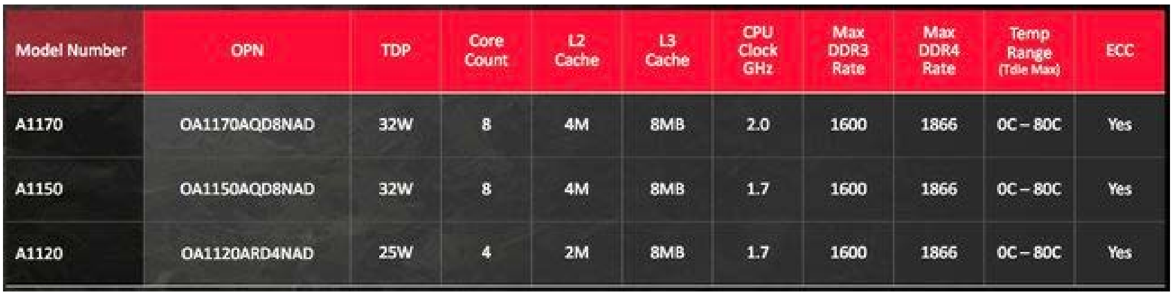
ARM processors follow two separate paths, the mobile and the HPC path. In both cases the processors now ship the eighth version of the ARM cores with 64-bit architecture. Several companies are preparing (or have prepared) their own SoC: Apple, Nvidia, Cavium, AppliedMicro and a few others.

Taking a closer look at ARM chips that aim to join the HPC league, we find AMD, AppliedMicro and Cavium.

AMD

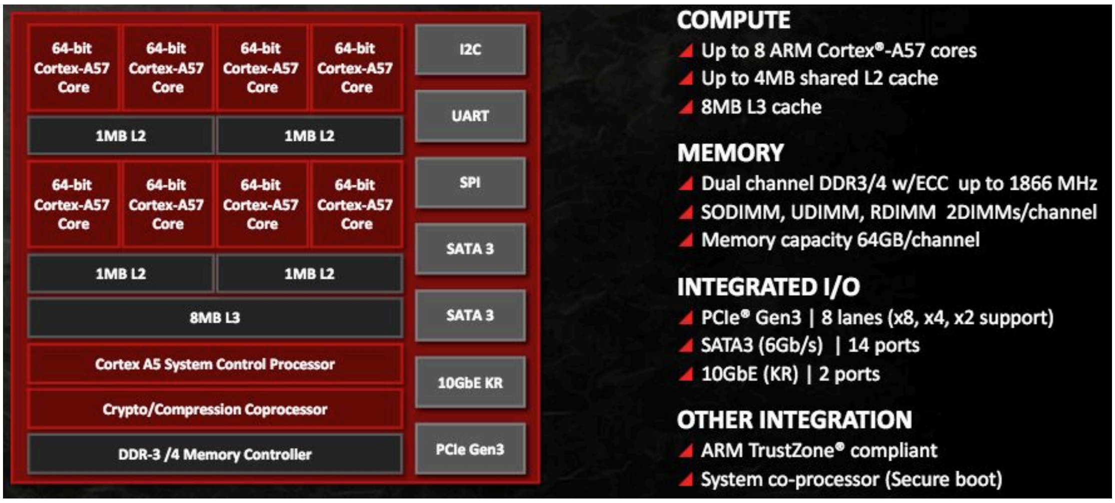
AMD is shipping the Opteron A1100, which is a 64-bit ARM-based server chip and has the following specifications:

AMD presents this new chip with two main points of interest: the 64-bit and ECC RAM (which hasn't been present on the previous 32bit ARM processors). It also adds a system control processor: a little Cortex-A5 core that serves as a remote management tool, integrated into the main System on a chip (SoC). This A5 includes a co-processor for accelerating cryptographic algorithms. As AMD stated, this new processor has a very low consumption overall.



Model Number	OPN	TDP	Core Count	L2 Cache	L3 Cache	CPU Clock GHz	Max DDR3 Rate	Max DDR4 Rate	Temp Range (Tdie Max)	ECC
A1170	OA1170AQD8NAD	32W	8	4M	8MB	2.0	1600	1866	0C – 80C	Yes
A1150	OA1150AQD8NAD	32W	8	4M	8MB	1.7	1600	1866	0C – 80C	Yes
A1120	OA1120ARD4NAD	25W	4	2M	8MB	1.7	1600	1866	0C – 80C	Yes

Figure 11: AMD A1000 series processor models



COMPUTE
- Up to 8 ARM Cortex®-A57 cores
- Up to 4MB shared L2 cache
- 8MB L3 cache

MEMORY
- Dual channel DDR3/4 w/ECC up to 1866 MHz
- SODIMM, UDIMM, RDIMM 2DIMMs/channel
- Memory capacity 64GB/channel

INTEGRATED I/O
- PCIe® Gen3 | 8 lanes (x8, x4, x2 support)
- SATA3 (6Gb/s) | 14 ports
- 10GbE (KR) | 2 ports

OTHER INTEGRATION
- ARM TrustZone® compliant
- System co-processor (Secure boot)

Figure 12: AMD A1000 series architecture

AppliedMicro

AppliedMicro (APM from now on) is targeting the ARM-HPC ecosystem with X-Gene 3. Built in 16nm, it will feature 32 cores operating at speeds up to 3.0 GHz, eight DDR4-2667 memory channels and 42 PCI-E 3.0 lanes. Performance is expected to be four-to-six times that of the previous X-Gene.

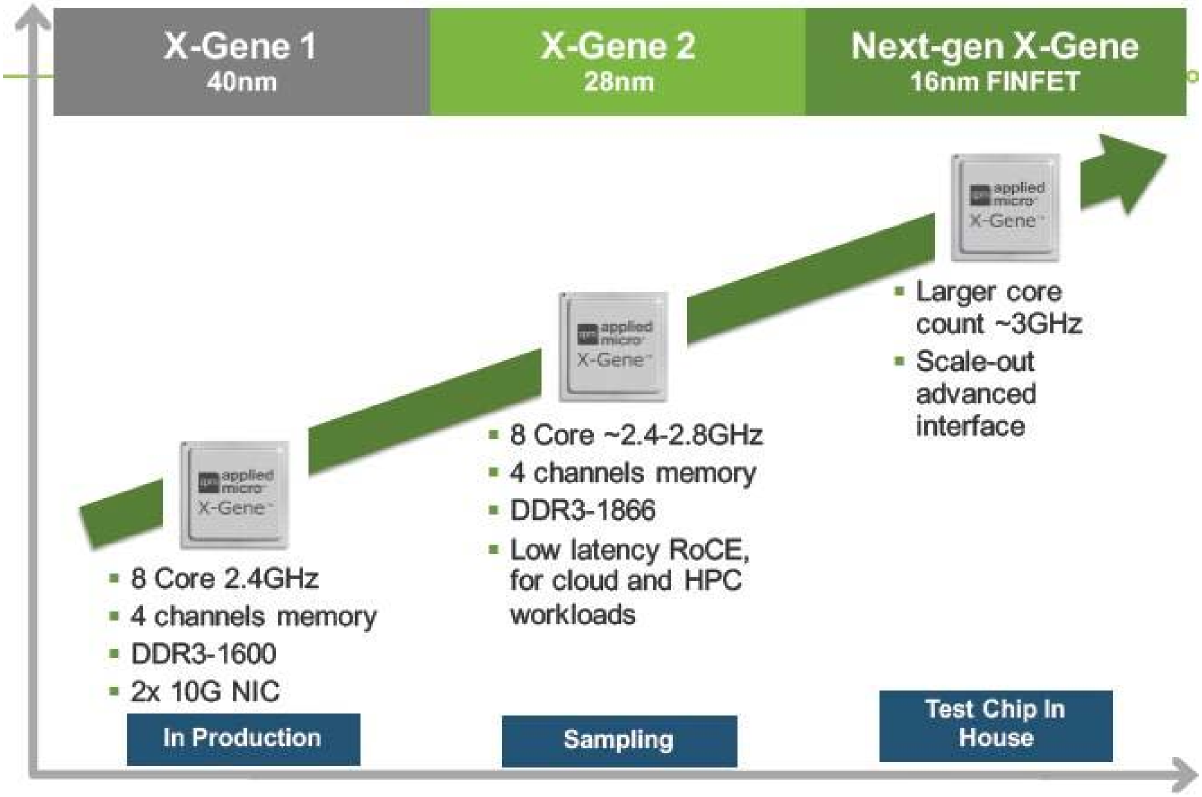


Figure 13: APM X-Gene Roadmap

Cavium

Cavium has the ThunderX line of ARM processors: up to 48 cores with up to 2.5 Ghz core frequency, DDR3/4 supporting 2400 Mhz with a maximum 1TB of memory in a dual socket configuration and 10/40/100 GbE connectivity.

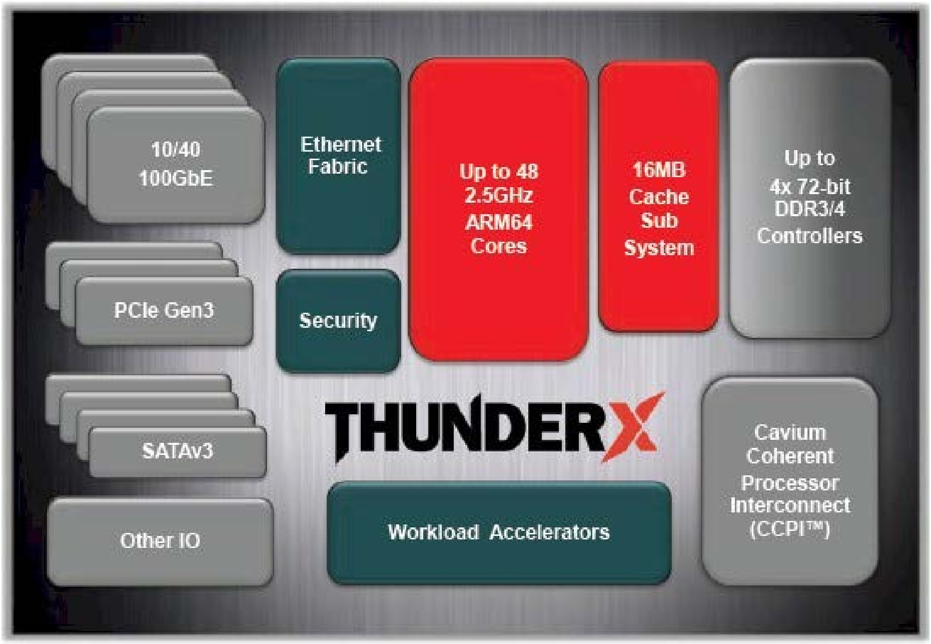

10/40
100GbE
PCIe Gen3
SATAv3
Other IO
Ethernet Fabric
Security
Up to 48 2.5GHz ARM64 Cores
16MB Cache Sub System
Up to 4x 72-bit DDR3/4 Controllers
THUNDERX
Workload Accelerators
Cavium Coherent Processor Interconnect (CCPI™)

Figure 14: ThunderX high-level CPU architecture

Mont-Blanc project and the ARM ecosystem

Mont-Blanc FP7 and H2020 projects are listed in chapter 6. Here is a specific subsection emphasising their relation to the current ARM ecosystem section.

Several efforts have been made by the HPC community to encourage the use of low-powered SoC embedded processors in large-scale HPC systems. These initiatives have successfully demonstrated the use of ARM SoCs in HPC systems, but there is still a need to analyze the viability of these systems for HPC platforms before a case can be made for Exascale computation. The Mont-Blanc project aims at providing an alternative HPC system solution based on the current commodity technology: mobile chips.

The major shortcomings of current ARM-HPC evaluation initiatives include a lack of detailed insights about performance levels on distributed multicore systems and performance levels for benchmarking in large-scale HPC applications. Additional contributing factor is the lack of a mature software stack needed to achieve the maximum performance available.

The Mont-Blanc project started in 2011 with the idea of leveraging mobile technology for performing scientific computing. The project designed, built and set-up a 1080-node HPC cluster made of Samsung Exynos 5 Dual processors. The Mont-Blanc project established the following goals: to design and deploy a sufficiently large HPC prototype system based on the current mobile commodity technology; to port and optimize software stack and enable its use for HPC.

After that, in 2013 the Mont-Blanc phase 2 started, with the objective to keep exploring the ARM 64-bit architecture and focusing on topics like fault tolerance and resiliency, performance analysis tools and advanced programming models (OmpSs). The third phase began in October 2015, within H2020, more focused on defining the architecture of a balanced pre-exascale HPC compute node. The coordination has been moved from BSC to Atos/Bull (France), mainly because there is a strong attempt to productize the system developed in the research project in the near future.

Given all these points, the fact is that ARM processors do not seem to be ready for production HPC, yet. The idea of low power consumption and higher density has lost most of its original strength, as stated by the direction of the newest processors that will reach the market in the following years. But still, the lack of a complete and fully operational software stack keeps the ARM from being a dangerous rival to Intel (or even AMD).

3.1.4 *POWER - OpenPOWER*

Mid 2014 IBM has released its latest generation of processors based on the POWER ISA, namely POWER8 [38]. The processors are available as single-chip modules (SCM) and dual-chip modules (DCM) with up to 10 and 2x6 cores, respectively. In this context relevant features of the POWER8 architecture include the following:

- High number of hardware threads per core, which was increased from 4 to 8 from POWER7 to POWER8, respectively.
- Support of high clock speeds in the range of 2.5 to 5 GHz [39].
- Relatively large size of caches, e.g. 8 MByte L3 cache per core.
- Outstanding memory bandwidth obtained by attaching memory to memory buffers (called Centaur) that are connected via bi-directional high-bandwidth links to the processor. For a socket with all 8 memory channels populated the nominal bi-section bandwidth between processor and the Centaur chips is 192 GByte/s when operating the links at a data rate of 8 Gbit/s.

- With 2, 128-bit wide SIMD units the peak double-precision floating-point operation throughput is 8 Flop/cycle per core. A 10-core POWER8 processor running at 3 GHz thus has a peak performance of 240 GFlop/s.

Before releasing POWER8, IBM announced in 2013 together with Google, Mellanox, NVIDIA and Tyan the OpenPOWER initiative. This initiative resulted in the setup of the OpenPOWER Foundation [40], which describes itself as an open technical membership organization. Several suppliers started to introduce products using the IBM POWER processor. At least one supplier announced its intention to provide customized processors based on POWER [41]. Furthermore, the OpenPOWER Foundation is providing a growing set of standards related to systems based on POWER processors. The goal is to establish an ecosystem where it becomes easier for suppliers to provide new products. This initiative thus is likely to result in a larger number of suppliers providing products based on POWER processors, including products that are relevant for HPC.

The following technical developments are of immediate interest for future HPC systems:

- NVIDIA and IBM announced that future versions of the POWER processor will support NVLink [42]. This will result in a significant increase of the bandwidth available for communicating data between processor and GPU and thus a tighter integration of both.
- Starting with POWER8 a new interface for coherent attachment of accelerators called CAPI is supported (for details, see [43]). Mellanox announced that it will use CAPI to attach their EDR Infiniband network interfaces claiming that it helps to reduce overheads [44]. CAPI will also facilitate attachment of FPGAs that are becoming a potentially interesting option for accelerating HPC work-loads.

The next generation of IBM POWER processors called POWER9 will be based on the 14 nm process and will be used for the next generation of supercomputers at Oak Ridge National Lab (ORNL) and Lawrence Livermore National Lab (LLNL). Initial deployments will include the CORAL systems which are co-developed with Mellanox and NVidia, which have aligned quite closely with IBM. The target delivery date for the ORNL system is 2017 [45].

3.1.5 *SPARC*

SPARC is still part of HPC landscape, with a unique integrator: FUJITSU.

In the latest Top500 list, this processor architecture is 5% of performance share and best ranks #4 on Top500 November 2015 list and #1 in Graph500 with K-computer. While K-computer is built with 88128 SPARC64 VIIIfx, 65 nm technology, 128 Gflops each, the following Fujitsu PRIME HPC embeds:

- SPARC64 IXfx 40 nm technology, 236 Gflops
- SPAR64 XIfx 20 nm technology, 1 Tflops

They key characteristics of SPARC64 XIfx are:

- Rich micro architecture: 32 computing +2 assistant cores
- Rich 256 bit wide SIMD function
- Interconnect on chip
- Support of HMC (Hybrid Memory Cube)

Fujitsu announced a 100 Pflops-capable system for Post FX-10 generation. Despite a lithography which is not at the state of the art (20nm vs. 14nm) for the latest HPC processor, the SPARC64 architecture, addressing efficiency on versatile operations, provides efficient HPC solutions and is part of the competitors for the race toward exascale.

3.1.6 *FPGA*

This year acquisition of Altera by Intel may result in hybrid devices that host a FPGA part next to one or more x86 cores. In the past the FPGA solutions proved to be not popular for HPC purposes due to high development effort required to make applications use this kind of solutions. With Intel plans for having a mathematical library that implements standard functions on FPGA chip this situation may change but so far there are no concrete declarations regarding the availability of expected performance.

3.2 Highly parallel components and compute engines

It is understood that the working frequency of silicon transistors used in modern CPU could not be increased any longer due to energy and thermal constraints. This imply that the performance of single functional unit does not increase any longer as well, e.g. the floating point unit deliver and will continue to deliver roughly one or few Gigaflops per second. As a consequence in order to continue to increase the performance of HPC system one obvious possibility is to increase the number of functional units, using design that maximize the throughput per watt to keep the overall power consumption as low as possible.

Today two mainstream architectural solutions are available following these principles. One is based on SoC design with many simple core featuring large vector units able to perform up to 32 Flops per cycle, while the other is based on a hybrid node design where a co-processor is paired with a traditional CPU to accelerate a set of instructions.

Both approaches allow multi-Teraflops per nodes as of today, and tomorrow an Exaflop/s with less than 100K nodes.

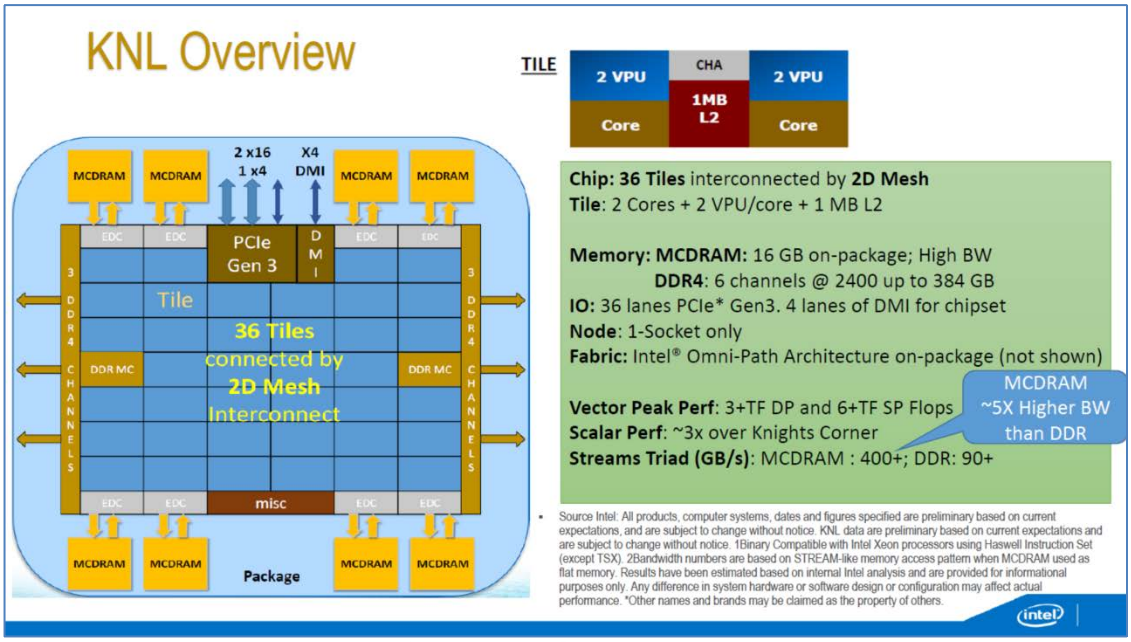
Here we report of few solutions interesting for designing HPC systems. In particular we report about Intel Knight Landing chip and new ARM based chip, of the first kind, NVidia GPU, other GPU solutions and FPGA.

3.2.1 *Intel Xeon Phi*

Intel Many Integrated Core Architecture or Intel MIC is a co-processor computer architecture developed by Intel incorporating earlier work on the Larrabee many core architecture, the Teraflops Research Chip multicore chip research project, and the Intel Single-chip Cloud Computer multicore microprocessor. At the International Supercomputing Conference (2012, Hamburg), Intel announced the branding of the processor product family as Intel Xeon Phi.

Prototype products code named Knights Ferry were announced and released to developers in 2010. The Knights Corner product was announced in 2011 and uses a 22 nm process. A second generation product codenamed Knights Landing (KNL) using a 14 nm process was announced in June 2013.

At the Supercomputing Conference'15, Intel has revealed some more details about its upcoming Knight's Landing platform. As the successor to Intel's existing Knights Corner (KNC, 1st generation Xeon Phi), Knights Landing makes the jump from using Intel's enhanced Pentium 1 (P54C) x86 cores to the company's modern Silvermont x86 cores, which currently lie at the heart of the Intel's Atom processors. These Silvermont cores are more capable than the older P54C cores and should significantly improve Intel's single threaded performance. All the while these cores are further modified to incorporate AVX units, allowing AVX-512F operations that provide the bulk Knights Landing's computing power and are a similarly potent upgrade over Knights Corner's more basic 512-bit SIMD units.



KNL Overview

Chip: 36 Tiles interconnected by **2D Mesh**
Tile: 2 Cores + 2 VPU/core + 1 MB L2

Memory: MCDRAM: 16 GB on-package; High BW
DDR4: 6 channels @ 2400 up to 384 GB
IO: 36 lanes PCIe* Gen3. 4 lanes of DMI for chipset
Node: 1-Socket only
Fabric: Intel® Omni-Path Architecture on-package (not shown)

Vector Peak Perf: 3+TF DP and 6+TF SP Flops
Scalar Perf: ~3x over Knights Corner
Streams Triad (GB/s): MCDRAM : 400+; DDR: 90+

- Source Intel: All products, computer systems, dates and figures specified are preliminary based on current expectations, and are subject to change without notice. KNL data are preliminary based on current expectations and are subject to change without notice. 1Binary Compatible with Intel Xeon processors using Haswell Instruction Set (except TSX). 2Bandwidth numbers are based on STREAM-like memory access pattern when MCDRAM used as flat memory. Results have been estimated based on internal Intel analysis and are provided for informational purposes only. Any difference in system hardware or software design or configuration may affect actual performance. *Other names and brands may be claimed as the property of others.

Figure 15: Intel KNL overview

Landing comes in many versions, depending on the number of cores and the packaging technology. Two distinctive features differentiate Knight Landing from its predecessors; the first one is that it is a standalone processor, like other processor of the Xeon family, even if a co-processor version available through a PCI card like GPUs, is in the roadmap of Intel. The second is that it is the first commercial processor featuring high bandwidth on package memory to mitigate the performance gap between cores and external DDR ram.

Intel is planning to offer Knights Landing processors containing up to 72 of these cores, with double precision floating point (FP64) performance expected to exceed 3 TFlop/s and over 6 TFlop/s in single precision.

Knights Corner already using a very wide (512-bit) GDDR5 memory bus, Intel is in need of an even faster memory technology to replace GDDR5 for Knights Landing. To accomplish this, Intel and Micron have teamed up to bring a variant of Hybrid Memory Cube (HMC) technology to Knights Landing. Called Multi-Channel DRAM (MCDRAM), Intel and Micron have taken HMC and replaced the standard memory interface with a custom interface better suited for Knights Landing. The end result is a memory technology that can scale up to 16GB of RAM while offering up to 500GB/sec of memory bandwidth (nearly 50% more than Knights Corner's GDDR5), with Micron providing the MCDRAM modules.

Contrary to the 1st generation KNC which was only available as a PCIe add-in coprocessor card, KNL can either be the main processor on a compute node, or as a co-processor in a PCIe slot. Typically Xeon Phi has an internal OS to access the hardware, but with this new model it eliminates the need for a host node – placing a KNL in a socket will give it access to both 16GB of high speed memory (the MCDRAM) as well as six memory channels for up to 384GB of DDR4, at the expense of the Intel Omni-Path controller. The KNL will also have

36 PCIe lanes which can host two more KNC co-processor cards and another four for other purposes.

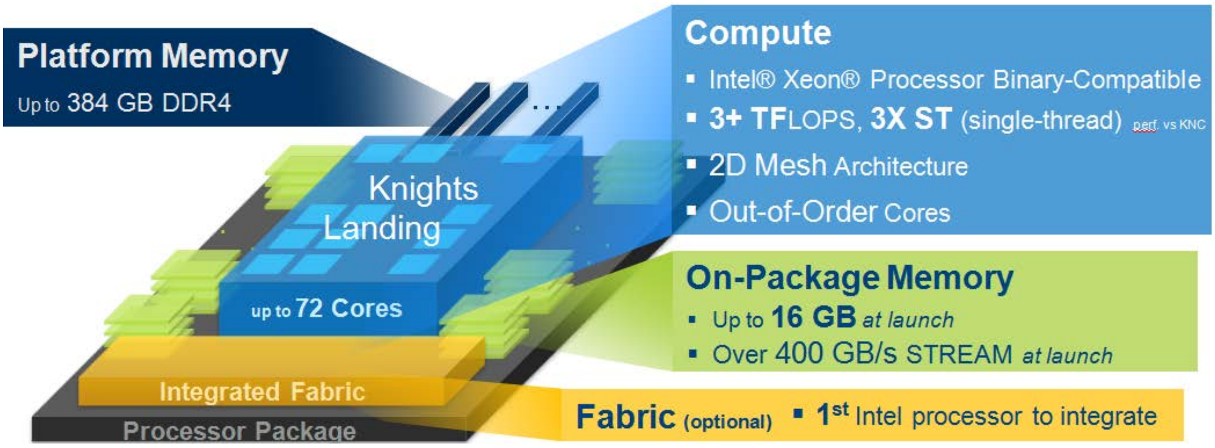


Figure 16: High Level picture of Knight Landing package

Knights Landing is planned for general availability in 2H2016. A version of KNL, known as KNL-F will integrate in the package an Omnipath network adapter still connected to the chip through PCI bus. KNL is expected to become general available in Q3 2016.

Figure 16 gives a high level picture of the Knights Landing package with high bandwidth 3D stacked memory and integrated fabric.

Two large supercomputing centres in US already announce systems that will feature KNL processors: Cori Supercomputer at NERSC, Trinity Supercomputer at NNSA, whereas the DOE and Argonne awarded Intel contracts for two systems (Theta and Aurora) as a part of the CORAL programme.

In Europe, CINECA's new Tier-0 system will be based mainly on Knight Landing Processors, integrated by Lenovo, but using Intel AdamPass platform and Intel Omnipath interconnect. CEA TERA1000 system will also encompass an important KNL configuration [47].

More public information about KNL is available on Intel site [46].

3.2.2 *NVIDIA GPUs*

The most successful example of processor co-processor hybrid design is those represented by the use of graphical processing unit (GPU) together with standard host CPU, to speed-up computation. This is not new, and is being used since almost 10 year now in high-end HPC production systems.

The leader of these particular markets is NVIDIA that dominates the Top500.

NVIDIA GPUs are in their 4th generation of GPU specifically designed for HPC and for accelerating floating point operations (Tesla, Fermi, Kepler, Maxwell). NVIDIA is now working on a new generation of these chips, code named Pascal and it successor Volta, with considerable new features to boost HPC applications performance. In particular the new design will integrate new busses and new memory to mitigate one of the main bottlenecks of current hybrid GPU design, the data movement required to offload computation to the GPU itself.

NVIDIA GPUs have already been announced to be the base for two new DOE supercomputers that will be built under the CORAL project, and integrated by IBM.

At this year NVIDIA revealed and confirmed two major bits about their next generation Pascal GPUs. The information includes details regarding process design, peak compute performance and even shared the same numbers for their Volta GPUs which are expected to hit the market in 2018 (2017 for HPC). The details confirm the rumours which we have been hearing since a few months that Pascal might be coming in market earlier next year (Figure 15).

Pascal will be the first Nvidia GPU to be built on TSMC's 16nm FinFET manufacturing process. The new process promises to be significantly more power efficient and denser than 28nm. This would enable Nvidia to build more complex and powerful GPUs all the while significantly improving power efficiency.

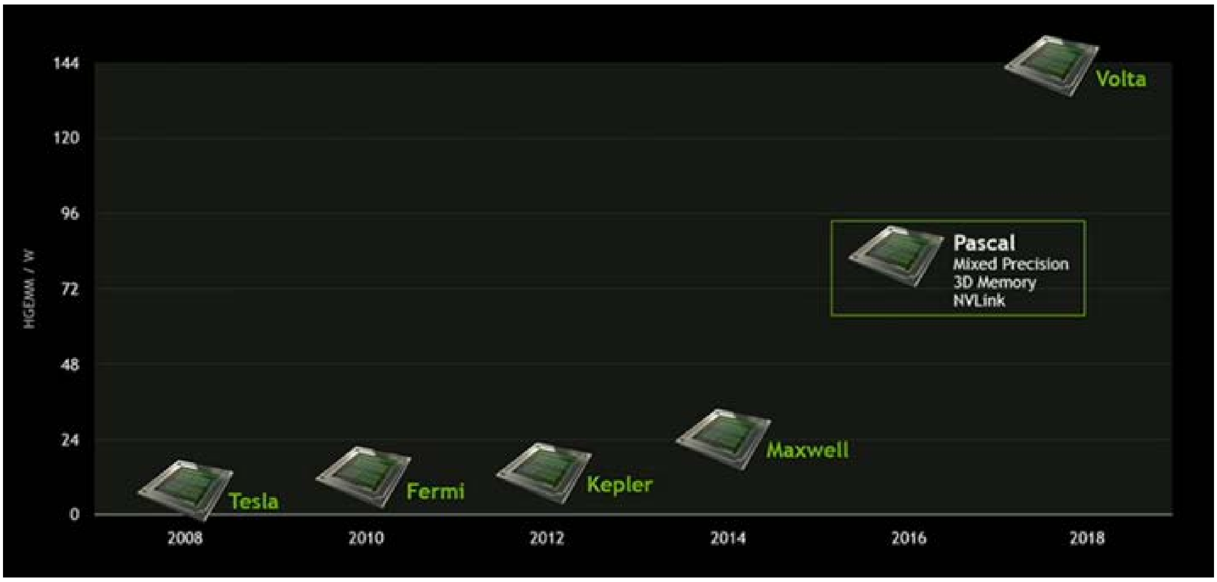
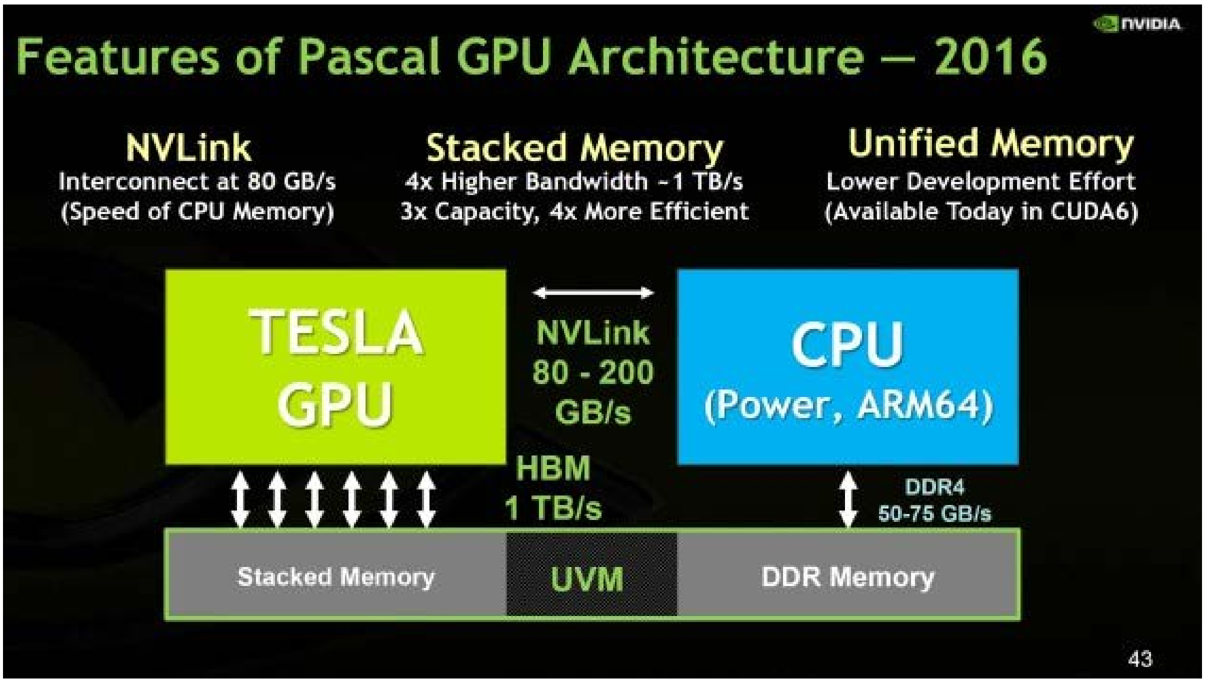


Figure 17: Public NVIDIA architecture roadmap.

Pascal (Figure 16) will be the first Nvidia GPU to feature the company's new NVLink technology and offers a UVM (Unified Virtual Memory) addressing inside a heterogeneous node. NVLink is a high-bandwidth, energy-efficient interconnect that enables ultra-fast communication between the CPU and GPU, and between GPUs. The technology allows data sharing at rates 5 to 12 times faster than the traditional PCIe Gen3 interconnect, resulting in dramatic speed-ups in application performance and creating a new breed of high-density, flexible servers for accelerated computing.



Features of Pascal GPU Architecture – 2016

NVLink
Interconnect at 80 GB/s
(Speed of CPU Memory)

Stacked Memory
4x Higher Bandwidth ~1 TB/s
3x Capacity, 4x More Efficient

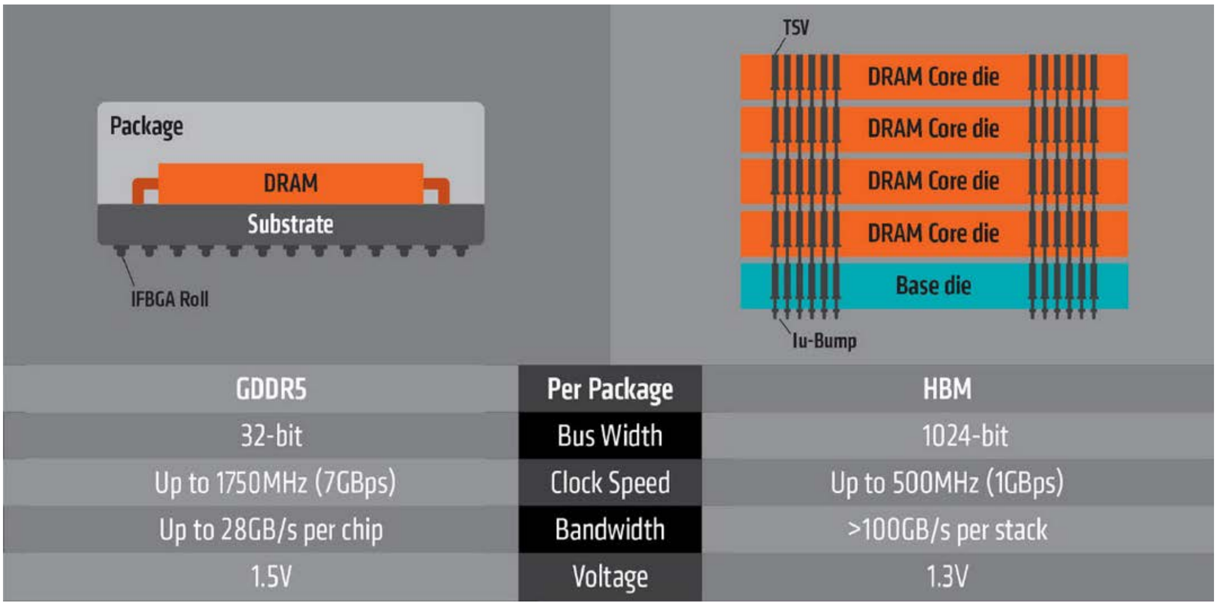
Unified Memory
Lower Development Effort
(Available Today in CUDA6)

Figure 18: NVIDIA Pascal architecture

The Pascal GP100 GPU has ~17 Billion transistors and around 32 GB of HBM2 based vRAM. With HBM2, NVIDIA gets the leverage to feature more memory than what's currently allocated on HBM1 cards (4 GB HBM on AMDs consumer line products: Fury X, Fury, Nano, Fury X2). With HBM2, NVIDIA gets access to denser chips that will result in cards with 16 GB and up to 32 GB of HBM memory across a 4096 bit memory interface. With 8Gb per DRAM die and 2 Gbps speed per pin, NVIDIA gets approximately 256 GB/s bandwidth per HBM2 stack. With four stacks in total, we will get 1 TB/s bandwidth on GP100 flagship Pascal three times that of the 980 Ti's 334GB/s.

On the compute side, Pascal is going to take the next incremental step with double precision performance rated over 4 TFlop/s, which is double of what's offered on the last generation FP64 enabled GPUs. As for single precision performance, we will see the Pascal GPUs breaking past the 10 TFlop/s barrier with ease. Pascal is still scheduled for a 2016 release with Volta coming after that date.

Figure 19 below gives the main characteristics of high bandwidth 3D stacked memory that will be available with Pascal card, as compared with today GDDR5 memory.



GDDR5	Per Package	HBM
32-bit	Bus Width	1024-bit
Up to 1750MHz (7GBps)	Clock Speed	Up to 500MHz (1GBps)
Up to 28GB/s per chip	Bandwidth	>100GB/s per stack
1.5V	Voltage	1.3V

Figure 19: Main characteristics of high bandwidth 3D stacked memory for Pascal

The next generation Pascal is characterized by two important architectural innovations: 3D high bandwidth memory and a new bus called nvlink. In particular nvlink will speedup memory transfers with respect to the current PCI bus, but require a CPU with a compatible bus to fully exploit it.

NVIDIA states that NVLink will use up to three times less energy to move data on the node at speeds 5-12 times conventional PCIe Gen3 x16.

NVLink enables fast communication between the CPU and the GPU, or between multiple GPUs. NVLink is a key building block in the compute node of Summit and Sierra supercomputers.

In its second generation (available in Volta architecture) the bus will be made compatible with IBM power processor to support CAPI (cache coherency protocol) and allowing for high speed data movement from host to GPU memory (Figure 19).

VOLTA GPU will feature NVLINK-2 for high-speed interconnect 80-200 GB/s and 3D Stacked Memory 4x Higher Bandwidth (~1 TB/s) and 3x Larger Capacity 4x More Energy Efficient per bit.

NVLink is a key technology in Summit's and Sierra's server node architecture, enabling IBM POWER CPUs and NVIDIA GPUs to access each other's memory fast and seamlessly. From a programmer's perspective, NVLink erases the visible distinctions of data separately attached to the CPU and the GPU by "merging" the memory systems of the CPU and the GPU with a high-speed interconnect. Because both CPU and GPU have their own memory controllers, the underlying memory systems can be optimized differently (the GPU's for bandwidth, the CPU's for latency) while still presenting as a unified memory system to both processors.

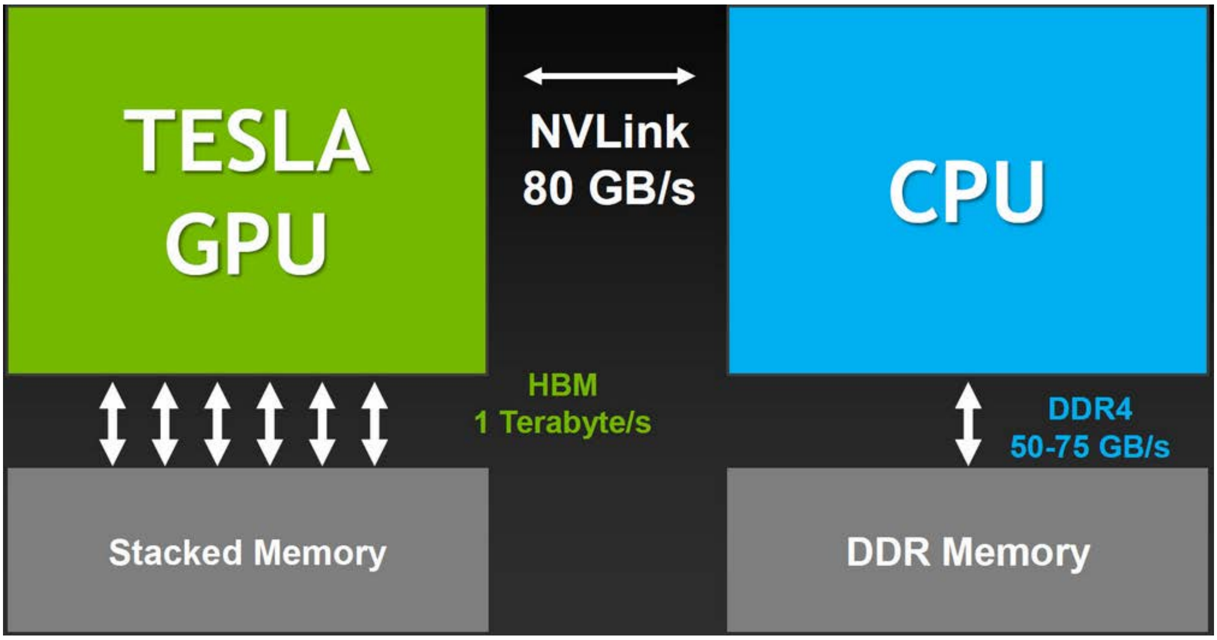


Figure 20: High level scheme of NVLink as used in the future IBM Power node

More information Pascal and Volta can be found in [145].

3.2.3 *PACS-G*

PACS-G (Processor Array for Continuum Simulation, Pax) is a processor that is built with exascale challenges in mind, in the context of Japan MEXT/RIKEN Flagship 2020 project [37]. The solution is designed to be implemented in the form of a PCI-connected accelerator card. The chip consists of multiple cores (PE), each with its own local memory that is used for computations. In addition to the local memory the PACS-G chip is to be equipped with global memory. This memory should be implemented as HBM or HBC 3D memory.

All PE units are internally connected by an in-chip communication network. Current plans assume usage of 4D mesh network for normal communication and dedicated network for reduction and broadcast operations.

Whole chip is planned to be used as a huge SIMD unit (all PE are executing the same instructions) with up to 4096 processing cores. Each PE should be able to execute 2FMA operations per cycle resulting in 16TFlops of theoretical peak performance per chip. The expected TDP of the chip is 250W which results in at least 50 Gflops/W

The supercomputer using this kind of solution is conceptually similar to currently utilized GPU clusters: node with general purpose processor equipped with one or more PACS-G accelerators. General purpose node acts as a gateway for storage while communication between accelerators is performed using dedicated inter-chip network.

This solution is designed with N-Body type problems in mind. XcalableMP and OpenACC are planned to be employed as programming platform. The release date for the chips is yet to be announced.

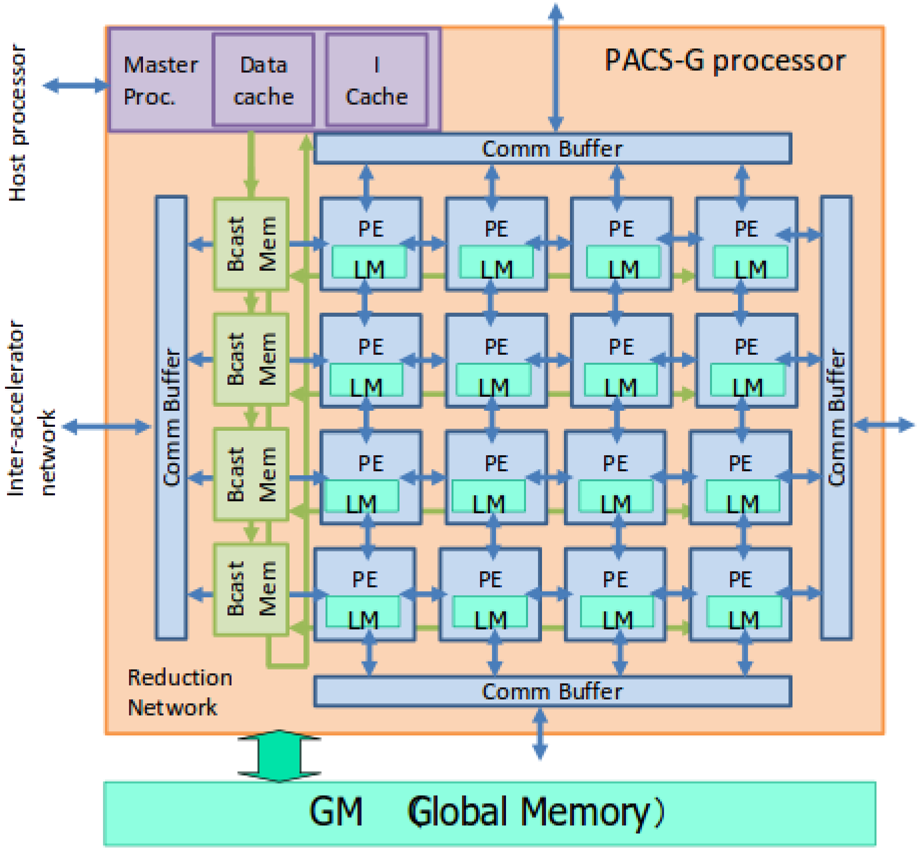


Figure 21: PACS-G processor

3.3 Memory and storage technologies

New memory technologies are expected to have a significant impact on the design of future high-end HPC architectures. While until recently DDR-SDRAM had been the dominating technology for realising off-chip memory layers in HPC architectures, the growing use of accelerator devices like GPUs and Xeon Phi resulted in having another memory layer based on high-bandwidth memory technologies. In future, a further change of even deepening of the memory hierarchy is expected when integration of non-volatile memory technologies is widely adopted. The introduction of expensive high-bandwidth memory technologies resulted in a stagnation of the aggregate memory capacity of high-end HPC systems at 1-2 PByte. With the integration of non-volatile memory technologies the memory capacity is expected to increase significantly.

3.3.1 *DDR-SDRAM*

With increasing support of DDR4 by different processor suppliers, this most recent generation of DDR-SDRAM technology is increasingly used. Its market share was expected to reach nearly 20% of the shipped DRAM bits [48]. With typical processor architectures like Intel Xeon E5-2600 v3 supporting data rates of up to 2,133 MT/s and 4 memory channels the nominal memory bandwidth is 68 GByte/s. Several new high-end HPC systems based on this processor technology and deployed in 2015 feature a DDR-SDRAM memory capacity of 64 GByte per socket.

Data rates and number of memory channels are expected to increase at a slow path. With 8-Gbit dies being more widely used, significantly higher memory capacity per node may become affordable.

3.3.2 *High-bandwidth memory technologies*

GDDR5 is currently the dominating high-bandwidth memory technologies. It is used for high-end GPUs as well as the first generation of Xeon Phi (Knights Corner). Currently data rates up to 6 Gbit/s are used, e.g. for NVIDIA K40. With a memory bus width of 384 bits the nominal memory bandwidth is 288 GByte/s. SK Hynix plans for data rates up to 8 Gbit/s in upcoming products [49]. Micron recently announced [53] support of an enhanced version of this technology, called GDDR5X, with data rates >13 Gbit/s. GDDR5X has been a JEDEC standard since November 2015 [54]. So far, no processing device supporting GDDR5X have been announced.

In the near future, processing devices based on the following new memory technologies will become available: HMC and HBM. Both technologies do foresee 3-dimensional stacking of DRAM dies.

HMC is a technology that is meanwhile standardized by the HMC Consortium, which is a working group made up of industry leaders [50]. Revision 2.1 of the standard was published in October 2015. The technology foresees a stack comprising a logic die and multiple DRAM dies. The logic die comprises multiple high-speed links to connect an HMC to another device, e.g. a processor. There are 4 links with up to 120 GByte/s per link (and direction). So far, only Fujitsu has announced a processor architecture based on HMC [51]. FPGA providers like Altera and Xilinx have shown test setup and made IP cores available (see, e.g., [52]). For the next generation of Xeon Phi (Knights Landing) a technology is used, which Intel calls Memory Cube Dynamic Random Access Memory (MCDRAM) and is believed to be largely based on HMC.

Unlike HMC, HBM only foresees an optional base logic die and a wide bus for connecting processing device and memory. Due to this width and the resulting large number of wires, HBM stacks have to be integrated into the package of the processing device. The current standard [63], also called HBM2, foresees up to 8 channels per stack each with a width of 128 bit. At a data rate of 2 Gbit/s the nominal bandwidth per stack becomes 256 GByte/s. AMD has brought first GPUs with HBM to the market, which however do not yet target the server market. NVIDIA is expected to release their next generation of HPC GPUs later in 2016, for which they use HBM2.

3.3.3 *Non-volatile memory technologies*

Non-volatile memory is a class of memory technologies that will become increasingly relevant for HPC in the near future. Except for non-volatility they also feature very high density, which significantly exceed the density of, e.g., DDR-SDRAM. Both features are important for using these technologies for storage devices.

Non-volatile memory can either be integrated in storage devices, which are attached through I/O interfaces like PCIe or SATA, or packaged in memory modules, i.e. Non-Volatile Dual In-line Memory Modules (NVDIMM). The former is optimized for capacity, while providing limited bandwidth, the latter provide a much smaller capacity over bandwidth ratio. At this point of time NVDIMM are hardly used for HPC, but this is expected to change with 3D XPoint (see below).

NAND flash

NAND flash memories do have a growing market with revenues exceeding USD 25 billion in 2012 [59]. Prices have been declining only slowly during recent years reaching about USD 0.25 per GByte in 2015 [59], which is significantly below DDR-SDRAM prices of about USD 8 per GByte. Recently, density of NAND flash memory could be increased using vertically oriented memory cells. The market share of this 3D NAND is growing but expected to stay below 25% in 2016 [55]. An important change for the integration of NAND flash devices in HPC architectures is the transition towards PCIe attached NAND flash storage devices [60]. This development is driven by the establishment of the NVMe interface (see below) as well as new device form factors, namely M.2 and U.2, targeting larger consumer markets.

A major limitation for a larger uptake of NAND flash for HPC architectures is the limited number of write operations a device can endure before being destroyed. Endurance is commonly described in terms of full Drive Writes Per Day (DWPD) for a certain warranty period, typically 3 or 5 years. High-end flash memory devices may support 5 DWPD and more over a period of 5 years, while for low-end devices this is typically 1 or less.

For specific purposes including check-pointing this endurance is expected to be good enough. For this reason NAND flash is increasingly often used in storage technologies like Cray DataWarp [58] or DDN IME [57] and others.

3D XPoint

3D XPoint (pronounced "cross point"), is an entirely new type of NVRAM (non-volatile memory) by Intel & Micron, with roughly 1,000 times the performance and 1,000 times the endurance of conventional NAND flash, while also being 10 times denser than conventional DRAM.

The first 3D XPoint memory chips will be available by the end of 2016, but there's no official timeline for commercialisation. Importantly, however, Intel and Micron say that 3D XPoint is "affordable", while other new memory technologies, such as phase-change memory, have so far proven too expensive to compete with gloriously cheap DRAM and NAND.

No transistors are involved into the memorization process, information is stored not using electrons but using a bulk property of the material in each memory cell.

In 3D XPoint, individual memory cells are read and written by simply applying a voltage to the appropriate word line and bit line. Unlike DRAM, there's no transistor governing each memory cell, which is one of the main reasons that 3D XPoint is (reportedly) much denser than DRAM.

Based on a three-dimensional arrangement of memory cells, allowing the cells to be addressed individually, 3D XPoint will be first available for new SSD devices and more significantly as a DIMM device.

3D XPoint based SSD sizes will be most probably 400, 800 and 1600GByte, whereas DIMMs will be available in 128, 256, 512Gbyte Memory Module (not confirmed).

3D XPoint DIMMs will be supported with so called apachepass interface for Skylake and Nighthill Intel processors. In particular Skylake has 6 memory channels, each one supporting 2 DIMMs one of which could host a NVRAM module, then 2 socket server (24 DIMM slots) can then be configured with up to 6TByte of NVRAM

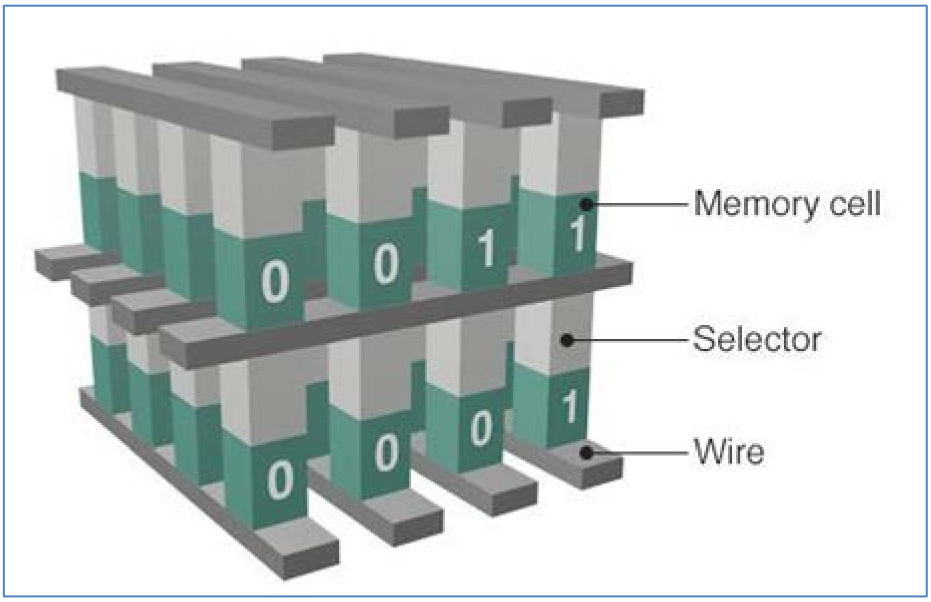


Figure 22: High level scheme of 3D XPoint chip

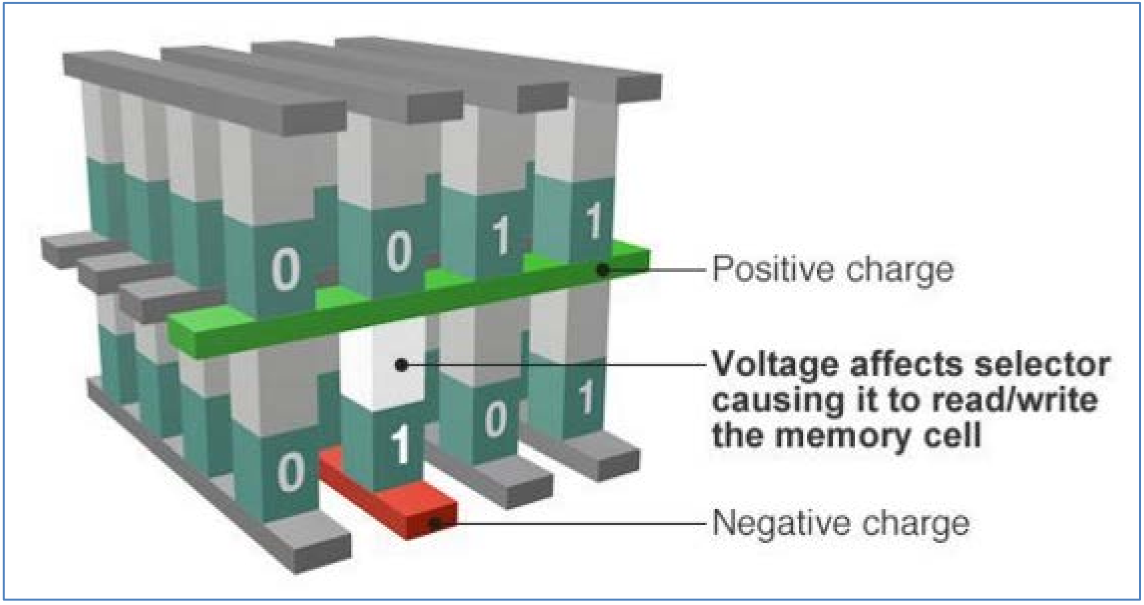

0
0
1
1
0
1
0
1
Positive charge
Voltage affects selector causing it to read/write the memory cell
Negative charge

Figure 23: High level representation of memory cell selection process

Non-volatile memory integration and programming models

NVM Express (NVMe) has become a de facto standard for host controllers to attach high-end non-volatile memory devices via PCIe. It allows to reduce processor load through function off-loading and to reach high bandwidth and I/O rates by enabling exploitation of hardware-level concurrency.

An interesting development for high-performance computing is the integration of non-volatile memory devices and network. One approach explored is to extend a local NVMe interface over networks supporting OpenFabrics RDMA protocols [56]. This would allow to aggregate non-volatile memory resources within a system to an addressable memory tier.

As of today, applications access non-volatile memory through a file system using POSIX or other I/O APIs. New programming models for accessing non-volatile memory have been standardized, see, e.g., [61], and libraries for managing persistent memory data structures are being developed [62].

Figure 5 NVM.PM.VOLUME and NVM.PM.FILE mode examples

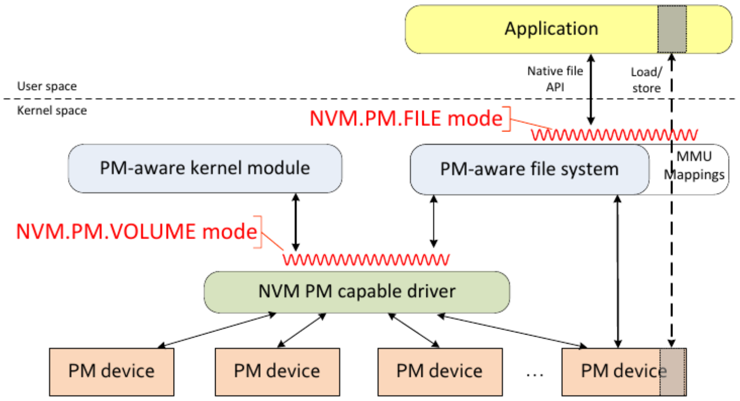


Figure 24: Future API towards non-volatile memory enabling direct load/store operations

3.4 Interconnect

High-speed interconnects are today mostly represented by Infiniband from Mellanox (Figure 26), but new technologies are emerging with the aim to get this portion of the market. Big vendors like Intel as well as smaller players like BULL will start to offer interesting alternatives in the following years.

Ethernet is still positioned number two, although for the last six years the number of systems is decreasing in favour of Infiniband with the exception of last year, where a lot of Chinese systems built with Ethernet (and often not designed for HPC) were benchmarked and introduced to TOP500 (see Figure 25).

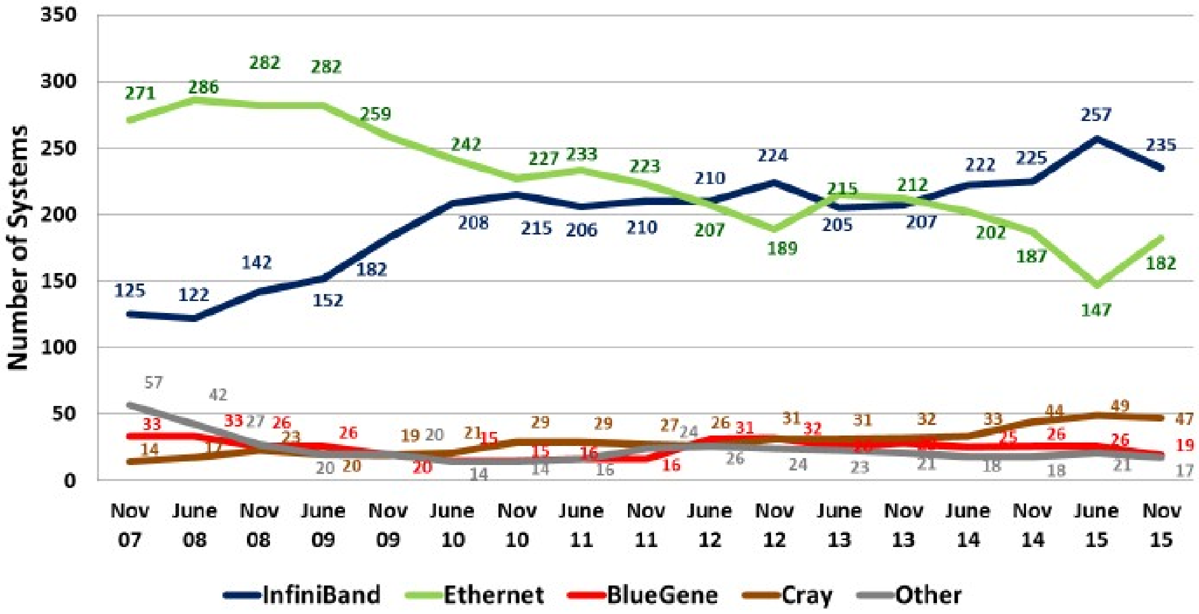


Figure 25: Trends of interconnect families (TOP500)

Special interconnects like Cray's Aries and Gemini or IBM's BlueGene/Q are the biggest representatives of custom and proprietary fabric technologies. Smallest in the system count, but important for the biggest systems are custom proprietary interconnects like the Tianhe-2 and Fujitsu TOFU [64].

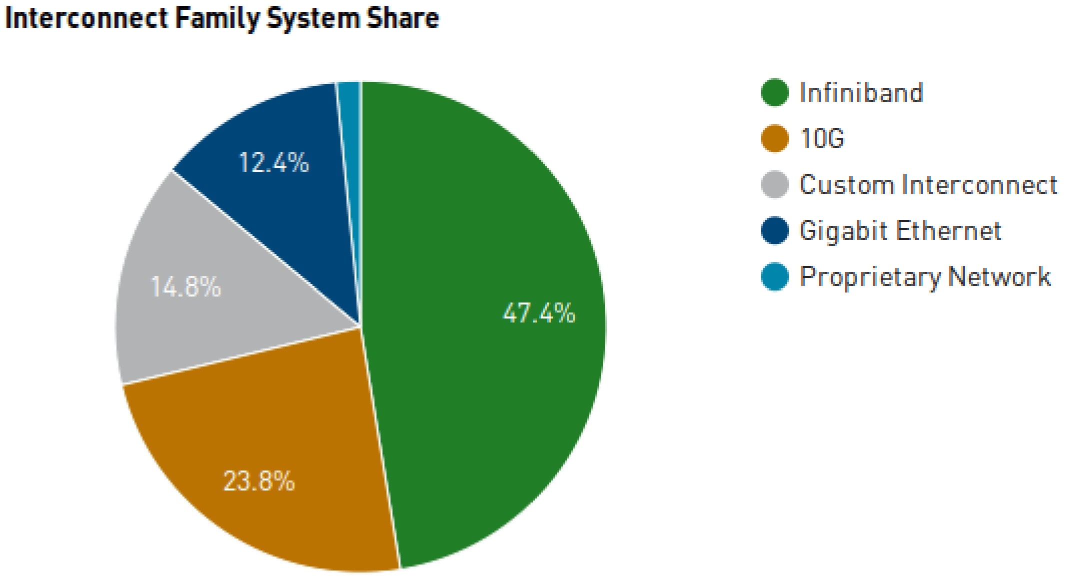


Figure 26: TOP500 list (November 2015) Interconnect family market share by systems

3.4.1 *Intel OmniPath & TrueScale*

Intel has already in the Interconnect portfolio two technologies, where the TrueScale fabric can be seen as an Infiniband QDR alternative or cheaper alternative to Infiniband FDR with the core technology acquired from QLogic (previously Voltaire) in 2012 and Ethernet as the HPC budget technology.

	TrueScale	OPA
SERDES Rate (Gb/s)	10	25.78
Peak bandwidth per port (Gb/s)	32	100
Messages rate (million/s)	35	160
Switch ports	36	48
Switch packet rate (million/s)	42	195
Switch latency (ns)	165-175	100-110

Table 3: TrueScale and OPA edge switch comparison

But Intel’s main focus should be on the new Intel Omni Path technology (OPA), which will be in 2016 in its first generation a real attempt to get serious market presence from Infiniband. OPA is based on the acquired intellectual property and patents from QLogic and Cray.

Table 3 shows the performance differences between TrueScale and OmiPath [65]. OPA is based on the acquired intellectual property and patents from QLogic and Cray and offers a full portfolio from adapters to director class switches and software stack (see Figure 27).

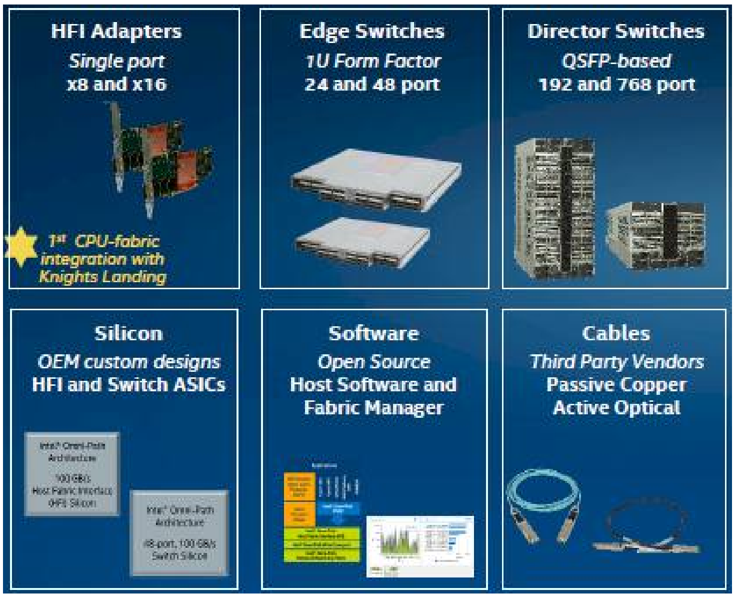


Figure 27: Intel OPA portfolio overview

Key features are:

- 100Gb/s on wire level speed with up to 2 ports
  Being on par with the Infiniband EDR version allowing more bandwidth, a standard PCIe gen 3 card (called HFI) will provide a single port, while the integrated version on KNL cards (and later on future CPUs) will provide a dual port version with up to 25GB/s bi-directional bandwidth;
- 48 ports switch ASIC;
  This allows the possibility to build bigger systems with less switches and cables resulting in either cost savings or having more money for the compute part of the system;
- Low latencies and High message rates;
  Intel promises to have 17% lower latencies and up to 7% higher messaging rates compared to Infiniband EDR. The switch ASIC should deliver 195 million MPI messages per second and the chip 160 million MPI messages per second in unidirectional messaging;
- Adaptive routing;
  Fabric manager monitors the congestion on the network and selects the best routing to avoid congested links ensuring the best throughput;
- Dispersive routing;
  This allows sending the data using multiple paths of the topology and better load balancing of the network;
- Traffic flow optimizations;
  This allows prioritizing certain message types, usually small MPI messages requiring small latency over big messages carrying data for network file system like LUSTRE;
- Packet integrity protection;
  Packet integrity is checked on every segment of the path (e. g. switch or HFI port) and in case of detected corruption, packet is re-sent on the affected segment only, eliminating the need of end-to-end retries.

3.4.2 *Ethernet*

Ethernet technology is today still present in HPC, where the biggest focus is on High Throughput Computing, Cloud computing and ScaleOut Computing where the extremely low latencies and high bandwidths of Infiniband are not needed.

The price/performance ratio is attractive with 10Gb/s or even more 1Gb/s per port speeds. But for 100Gb/s port level speed Ethernet is surprisingly not so interesting. Current comparison of 100Gb/s technologies is more favourable to Infiniband-like technologies (see Table 4 below) [66].

Vendor	Technology	Price per port (USD)
Mellanox	Switch-IB2 EDR	$333
Mellanox	Spectrum 100GbE	$590
Arrista/DELL/HPE	Ethernet 100GbE	$1000
Intel OPA	48 port non-managed switch	$386

Table 4: Price comparison of selected interconnects

A new Ethernet standard featuring the port speed of 25Gb/s has emerged last year to support a wider adoption of “cheap” Ethernet replacing the “slow” 10Gb/s version and to overcome the high price of the 40Gb/s and 100Gb/s versions (see Figure 28). This can be done using the same SFP connectors and cables as was used for 10Gb/s and 40Gb/s.

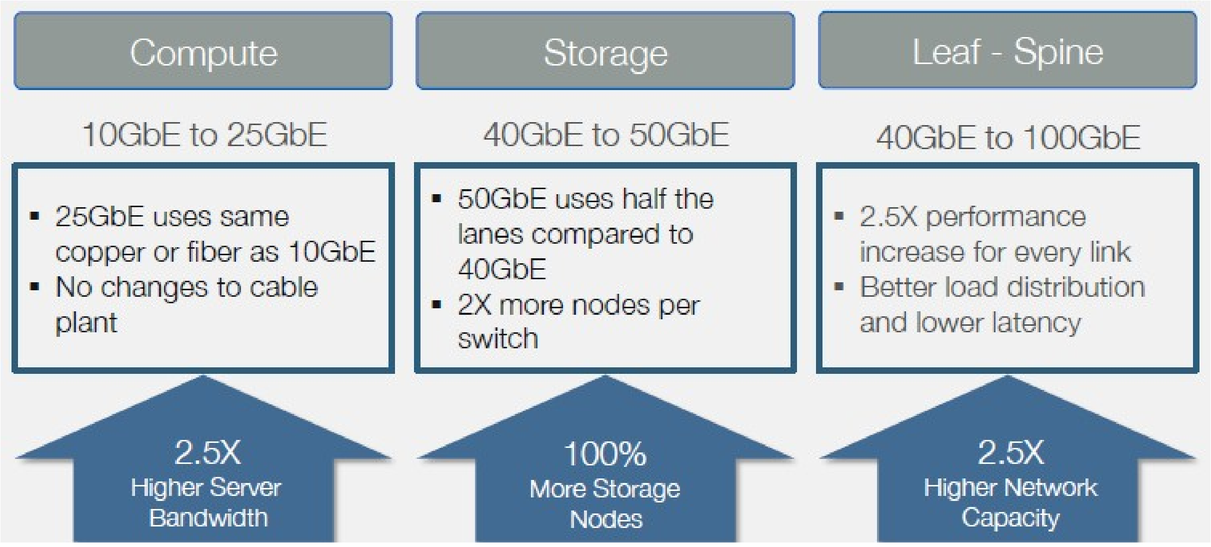

Compute
Storage
Leaf - Spine
10GbE to 25GbE
40GbE to 50GbE
40GbE to 100GbE
▪ 25GbE uses same copper or fiber as 10GbE
▪ No changes to cable plant
▪ 50GbE uses half the lanes compared to 40GbE
▪ 2X more nodes per switch
▪ 2.5X performance increase for every link
▪ Better load distribution and lower latency
2.5X Higher Server Bandwidth
100% More Storage Nodes
2.5X Higher Network Capacity

Figure 28: Overview of upgrade to 25GbE

Especially Mellanox with its Spectrum line of switches (Figure 28) supports the building of systems with edge switches and server cards running on 25Gb/s while the uplinks from edge switches to core switches will be on 100Gb/s links.

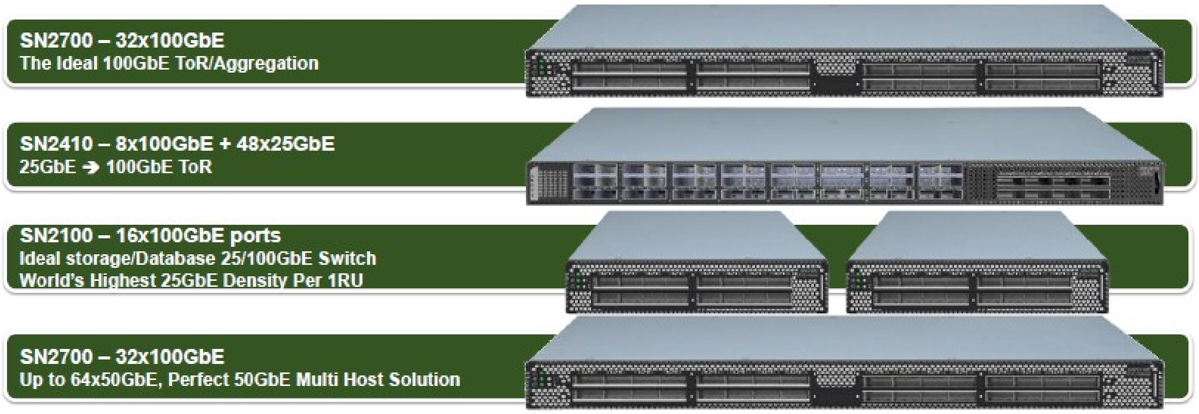
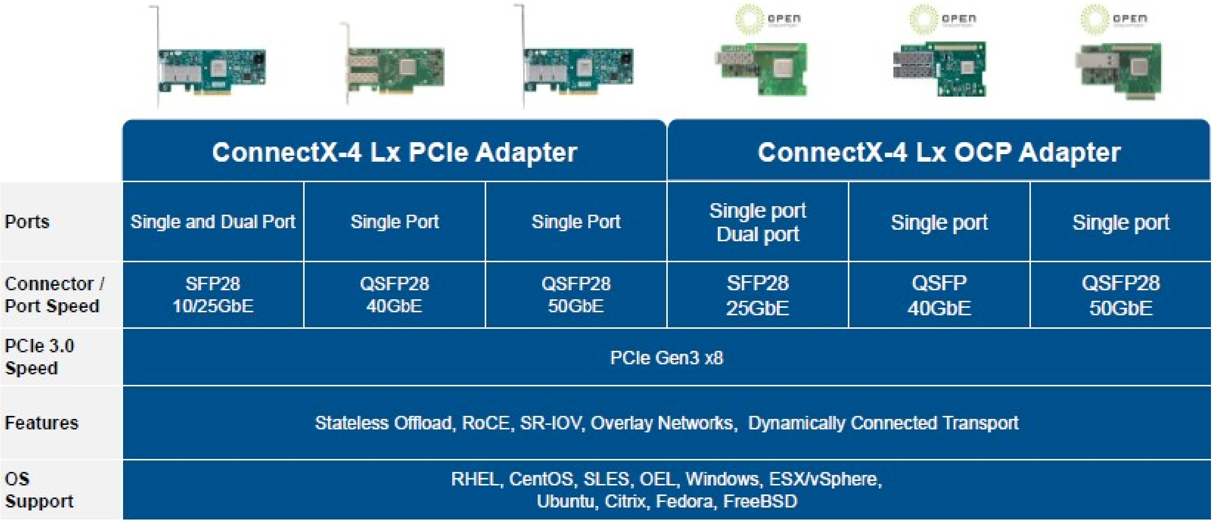

SN2700 – 32x100GbE
The Ideal 100GbE ToR/Aggregation
SN2410 – 8x100GbE + 48x25GbE
25GbE → 100GbE ToR
SN2100 – 16x100GbE ports
Ideal storage/Database 25/100GbE Switch
World's Highest 25GbE Density Per 1RU
SN2700 – 32x100GbE
Up to 64x50GbE, Perfect 50GbE Multi Host Solution

Figure 29: Mellanox Ethernet Switches portfolio

Ethernet from Mellanox can support up to 50Gb/s on a single port NIC (Figure 30) and switches with up to 64 ports running at up to 50Gb/s. The port to port latency on the Spectrum chip for switches has a 300ns port-to-port latency.



	ConnectX-4 Lx PCIe Adapter			ConnectX-4 Lx OCP Adapter		
Ports	Single and Dual Port	Single Port	Single Port	Single port Dual port	Single port	Single port
Connector / Port Speed	SFP28 10/25GbE	QSFP28 40GbE	QSFP28 50GbE	SFP28 25GbE	QSFP 40GbE	QSFP28 50GbE
PCIe 3.0 Speed	PCIe Gen3 x8					
Features	Stateless Offload, RoCE, SR-IOV, Overlay Networks, Dynamically Connected Transport					
OS Support	RHEL, CentOS, SLES, OEL, Windows, ESX/vSphere, Ubuntu, Citrix, Fedora, FreeBSD					

Figure 30: Mellanox Ethernet NIC overview

Other major players in Ethernet networks are Arista, DELL and HPE planning to use the new Broadcom ASIC code named Tomahawk.

Main features are:

- Support for up to 32 × 100 GbE, 64 × 40/50 GbE or even 128 × 25 GbE ports with an aggregate switching bandwidth of 3.2 Tbps
- Integrated low-power 25Ghz SerDes
- Configurable pipeline latency enabling sub-400 ns port-to-port operation
- Supports high-performance storage/RDMA protocols including RoCE and RoCEv2
- BroadView instrumentation provides switch- and network-level telemetry

3.4.3 *Bull BXI*

In 2016 BULL will start to offer a new interconnect called BXI – Bull eXascale Interconnect. At first this will be part of the HPC platform Sequana, but later it might be offered as a standalone product and thus become an alternative to both Infiniband and Intel Omni Path [67].

BXI will feature its own NICs as well as switches together with own software stack. Unlike the Intel OPA, which is partially based on Infiniband, BXI is not and thus will be incompatible with native HW and software (like IB verbs interface). On the other hand there are some similarities in design of OPA and BXI [68].

BXI will support in hardware SHMEM and PGAS implementations.

The software stack is built on Portals4 network API, developed by Sandia labs.

Standard protocols are supported in kernel driver, including IPoPtl (IP over Portals) for IPv4/IPv6 support and Portals4 Lustre Network Driver (Figure 31).

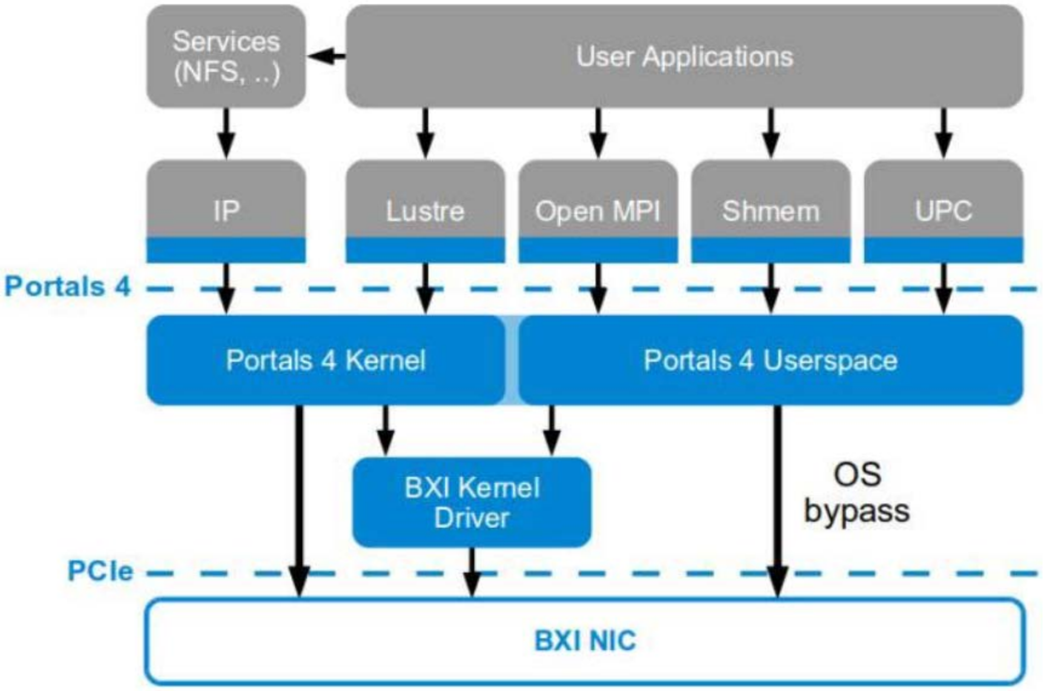


Figure 31: BXI software stack

Key features of BXI NICs (as depicted also on Figure 32) are:

- 100Gb/s on wire level speed
  Aligned with standard 100Gb/s Ethernet, so cabling and connectors are compatible. At 100Gb/s copper cables can be only 1.5m long. Standard PCIe format using PCIe gen 3 with up to 16 lanes.
- End-to-end error checking on NIC and Link level error checking
  Provides a mechanism for transient and permanent failures using message integrity (32-bit CRC with each message), message ordering for MPI messages (16bit sequence number), and message delivery for retransmission of lost or corrupted messages on every segment of the path.
- HW implementation for Portal4 communication primitives
  Allows overlapping computations and communication, supports MPI two-sided messaging and PGAS/MPI one sided messaging.
- OS bypass
  Allows the application to issue commands directly to the NIC bypassing the kernel and reducing latency.
- Offload operations

Logical to physical ID translation, MPI matching and Virtual to physical memory mapping without the OS and main CPU involvement. Collective operations using atomic and triggered operation units.

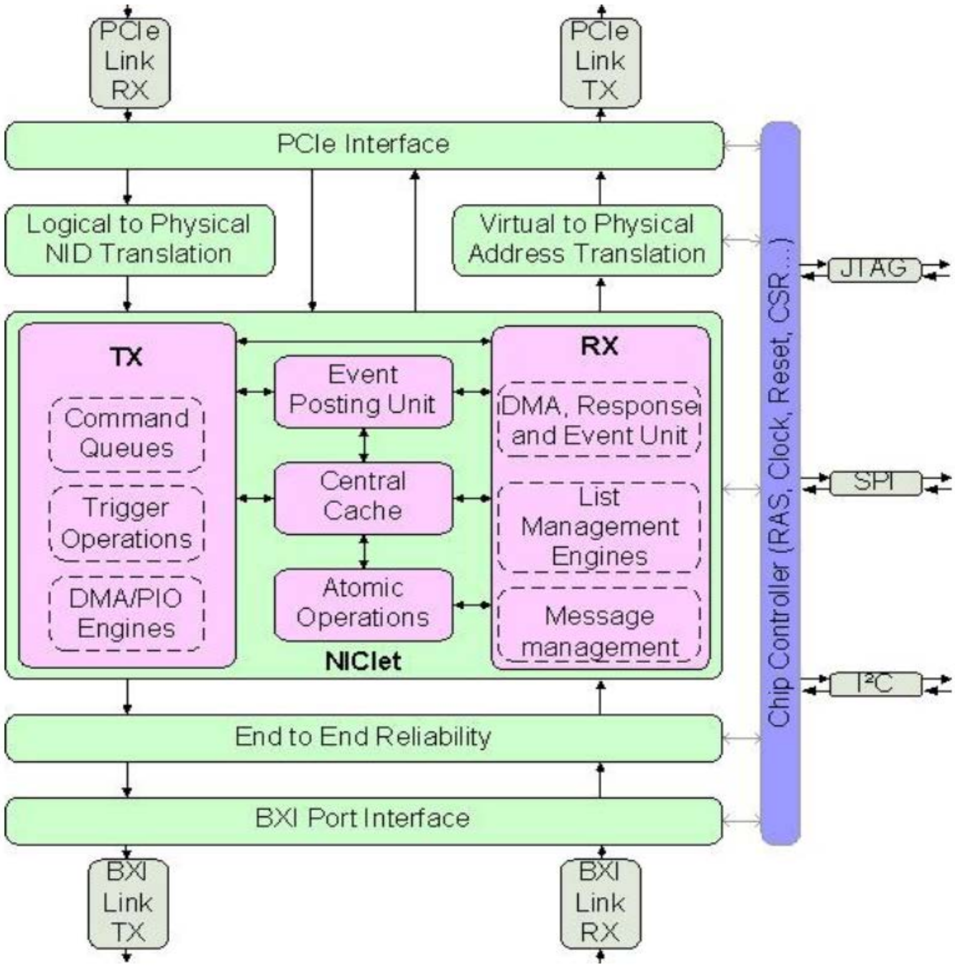


Figure 32: BXI NIC block diagram

Key features of BXI switch are:

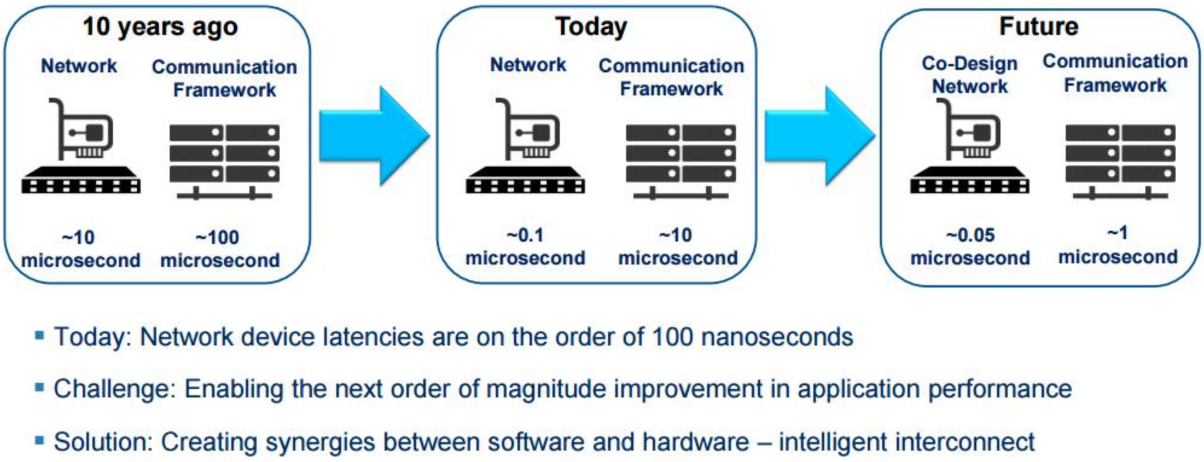
- 48 ports switch ASIC
  This allows up to 64,000 nodes big systems with full fat-tree topology.
- Multiple topologies
  For very large systems, topologies like Torus, Flattened Butterfly, Dragonfly and other can be used to lower the cost for the network. Of course standard tree topologies including fat-tree are supported.
- Adaptive routing
  Similar to dynamic routing to avoid congested links in the network.
- RAS features
  ASIC ECC, Out-of-band management link.

Performance estimations are:

- Latency below 1 micro second. Issue rate of 110 million uni-directional messages per second and 160 million bi-directional messages per second.
- Application level bandwidth up to 11GB/s.
- Time for a barrier or an 8-bytes allreduce on 32,000 nodes at 20µs.

3.4.4 *Infiniband*

Infiniband, a de-facto standard for HPC, is today dominated by Mellanox. The company reacted on the incoming new technologies from Intel and Bull and started with the new generation of EDR Infiniband to focus more on intelligent interconnects rather than raw performance numbers (Figure 33) [70]. The main focus is on acceleration of collective operations in MPI, SHMEM and PGAS by offloading them to the new switch ASIC called Switch-IB 2.



- Today: Network device latencies are on the order of 100 nanoseconds
- Challenge: Enabling the next order of magnitude improvement in application performance
- Solution: Creating synergies between software and hardware – intelligent interconnect

Figure 33: Mellanox future vision of intelligent network

Compared to the previous ASIC Switch-IB, the main characteristics are the same [71]:

- Port-to-port latency is around 86 nanoseconds
- 36 ports per ASIC
- 100Gb/s port bandwidth
- 7.2 TB/s aggregated bandwidth
- 7.02 billion messages per second
- Adaptive routing
- Multiple topologies (Fat Tree, 2D/3D Torus, 2D/3D mesh)
- RDMA and RoCE

The HCAs are use the latest generation of the chip ConnectX-4, with main features:

- EDR 100Gb/s InfiniBand or 100Gb/s Ethernet per port
- 1/10/20/25/40/50/56/100Gb/s speeds
- 150M messages/second
- Single and dual-port options available
- Virtual Protocol Interconnect (VPI)
- Power8 CAPI support
- CPU offloading of transport operations
- Application offloading
- Hardware offloads for NVGRE and VXLAN encapsulated traffic
- End-to-end QoS and congestion control
- Hardware-based I/O virtualization
- Ethernet encapsulation (EoIB)

The most interesting feature might be support for current and future Power architectures from IBM or OpenPower consortium. For that platform Infiniband might be the only reasonable HPC interconnect technology.

Mellanox plans to upgrade the Infiniband technology to HDR in 2017 with bandwidth increase up to 200GB/s.

3.4.5 *NUMAlink*

NUMAlink is the proprietary interconnect from SGI, currently used in the sixth generation of their coherent shared memory system SGI UV 3000 [72]. It provides high-bandwidth, low latency and coherency optimized functionality using the Intel QPI links to connect directly to the CPUs. In partitioned mode of UV, where the nodes run separate versions of OS, NUMAlink can support MPI, SHMEM and PGAS communication between the OS instances.

One NUMAlink ASIC provides a Global Register Unit (GRU) to extend the cache coherency across the whole system, Active Memory Unit for atomic memory operations, QPI interface with up to 32GB/s aggregated bandwidth and four NUMAlink ports with aggregated bandwidth of 40 GB/s to support multiple topologies based on 3D Enhanced Hyper-Cubes (3D EHC). The ASIC supports MPI Offload accelerating tasks like barriers and reductions.

A basic 3D EHC supports up to 8 nodes each with 2 CPUs. This is called an IRU (individual rack unit). The topology is shown on Figure 34.

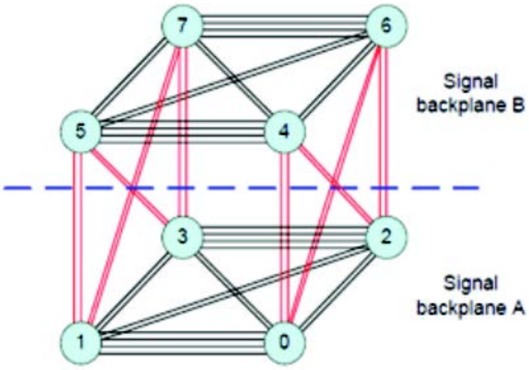


Figure 34: NUMAlink single IRU topology

Multiple IRUs (up to 16) are than connected using two level cross-bars between the individual 3D Enhanced Hyper-Cubes as shows Figure 35.

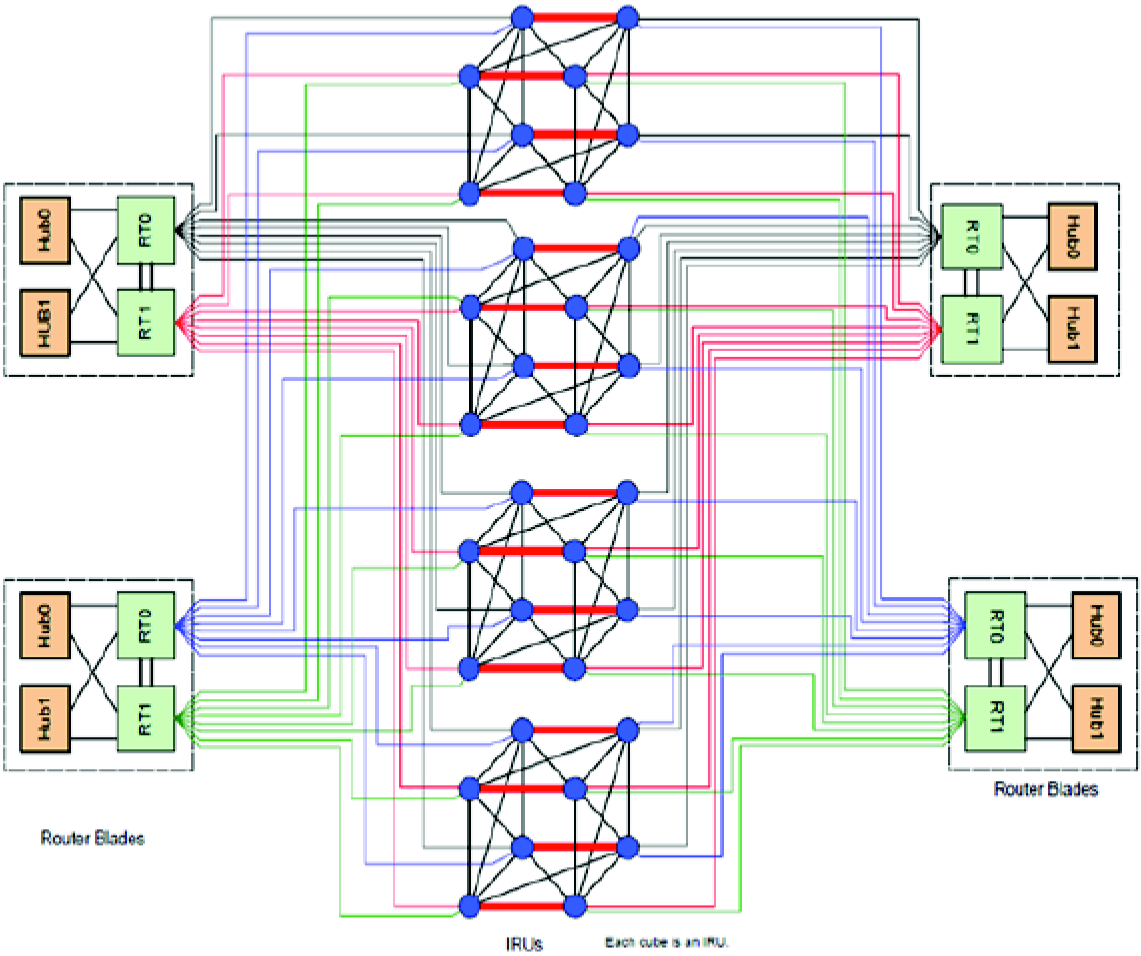


Figure 35: NUMAlink topology with 4 IRUs

With NUMAlink a very large cache-coherent systems can be built with sizes up to 256 CPUs and 64TB RAM.

3.4.6 *EXTOLL*

Extoll is the name of both a German company and their interconnect technology. The technology introduces network architecture for HPC with low latency, high message rates, low memory footprint, switch-less 3D Torus topology for up to 64,000 nodes configurations.

First important installation (still using the older version based on FPGA) is at the European Exascale project DEEP, used to connect the Xeon Phi coprocessors in the Booster part creating a 384 nodes system. The newer ASIC based version with immersive cooling is now used in smaller version called GREEN ICE Booster with 32 nodes [73].

Basic building block is the ASIC based card TOURMALET (Figure 36). The card connected via PCIe gen3 supports Remote Memory Access, HW Address Translation Unit (including TLB support), Adaptive Routing, CRC for network elements and ECC for memory and other features found in modern interconnects. The software stack supports Linux, OpenMPI, PGAS, GPU-to-GPU direct communication and provides low-level API for direct access to the hardware from applications. Each card has 7 ports with up to 100 Gb/s with switching capacity up to 1.75 Tb/s. The ports use a Samtec HDI6 connector [74].

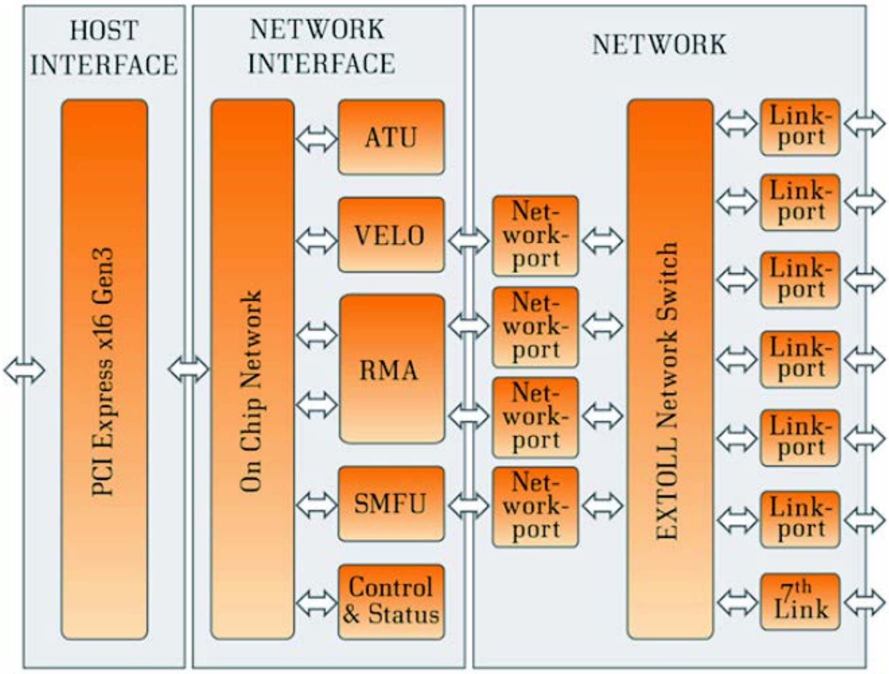


Figure 36: TOURMALET block diagram

To build a topology, the cards are connected using an Active Optical Cable (Figure 45) which can be from 2m to 10m long. Each cable has 12 bi-directional lanes, 24 OM3 fibres with up to 120Gb/s full-duplex bandwidth. Each cable consumes 2W at 3.3V. The cable exhibits less than 20ns of latency at 5m.

Figure 37: Custom Active Optical Cable from Extoll

3.4.7 *Cray ARIES*

Aries interconnect represents the mostly used custom interconnect (36 systems in last TOP500) and is used for all Cray XC systems. Cray will keep this interconnect for at least two following years [75].

The main building block is a SoC device connecting up to 4 compute nodes, each with two CPUs and providing a 48 port router/switch. This is shown on Figure 38 [76].

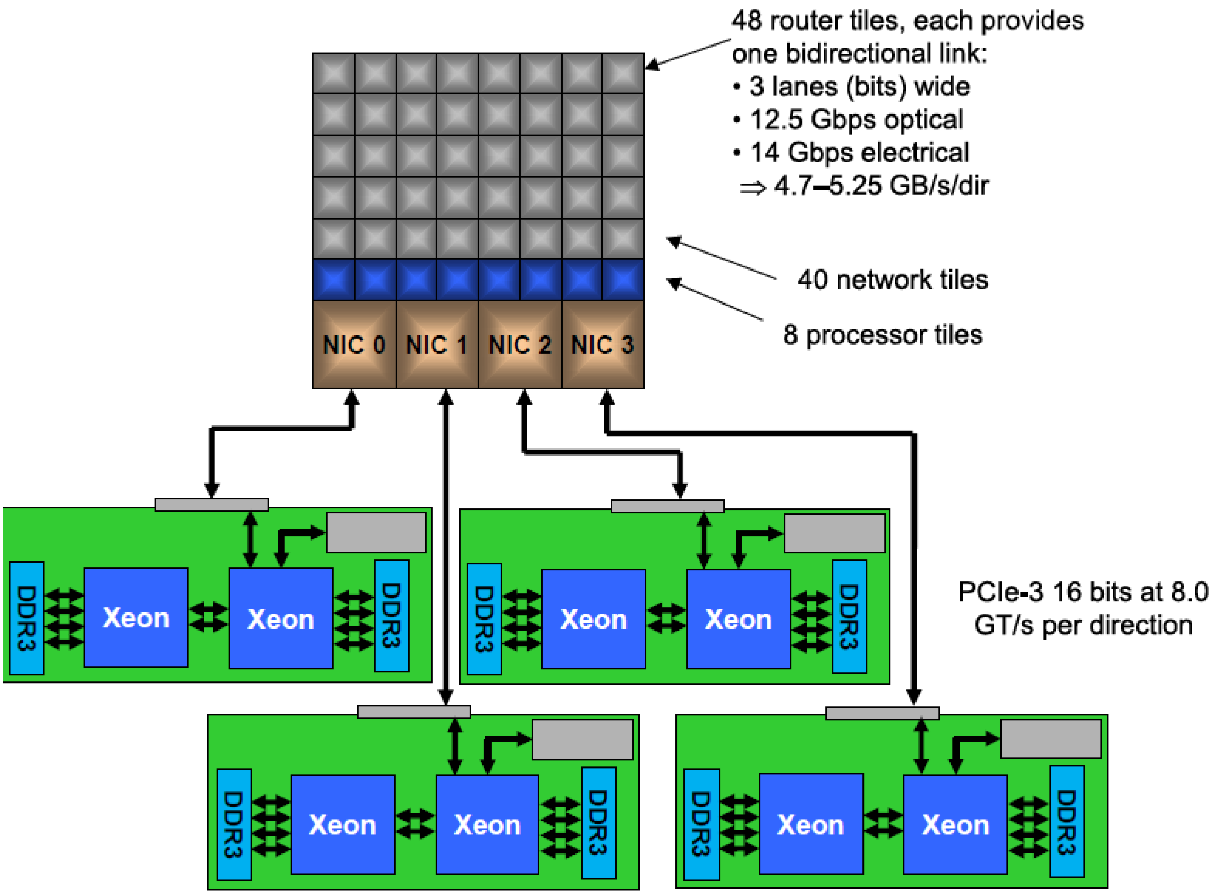


Figure 38: Aries system-on-a-chip device

Main features are:

- Dragonfly topology:
  Cost-effective and high-bandwidth topology. Up to 384 nodes connected just with metal wires and employing optical fibres only between such groups of nodes supporting systems up to 92,544 nodes;
- Adaptive routing to avoid congestion using load balancing;
- Remote memory access;
- RAS features including CRC and ECC on most components;
- Low latencies (1.3us for MPI on the same SoC, 100ns – 500ns between the SoCs);
- High-bandwidth (bi-directional up to 15GB/s and uni-directional up to 7.5GB/s);
- Complex software stack supporting Lustre, IP, MPI, SHMEM and PGAS, with the possibility to bypass OS for better application performance (see Figure 39).

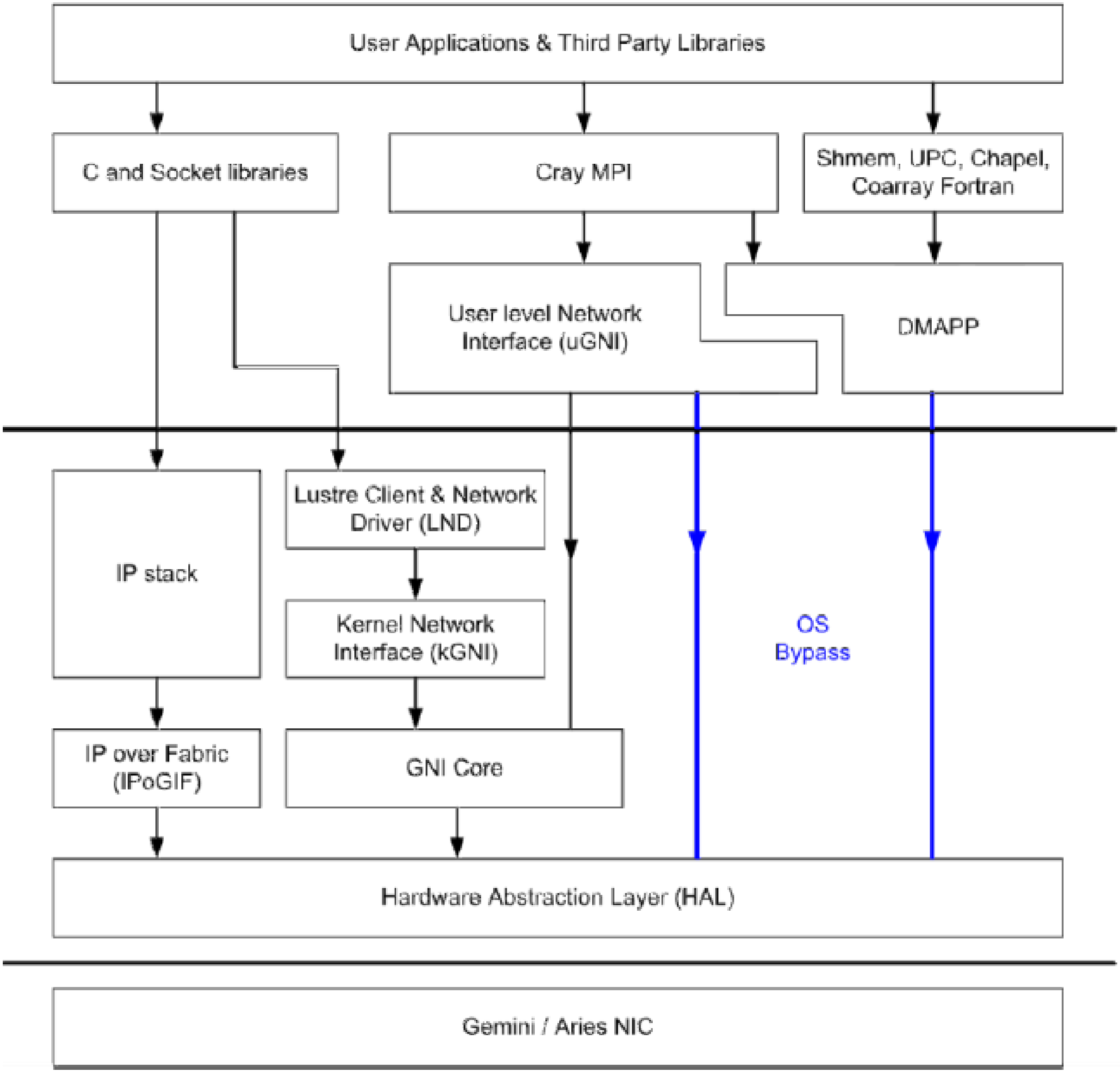
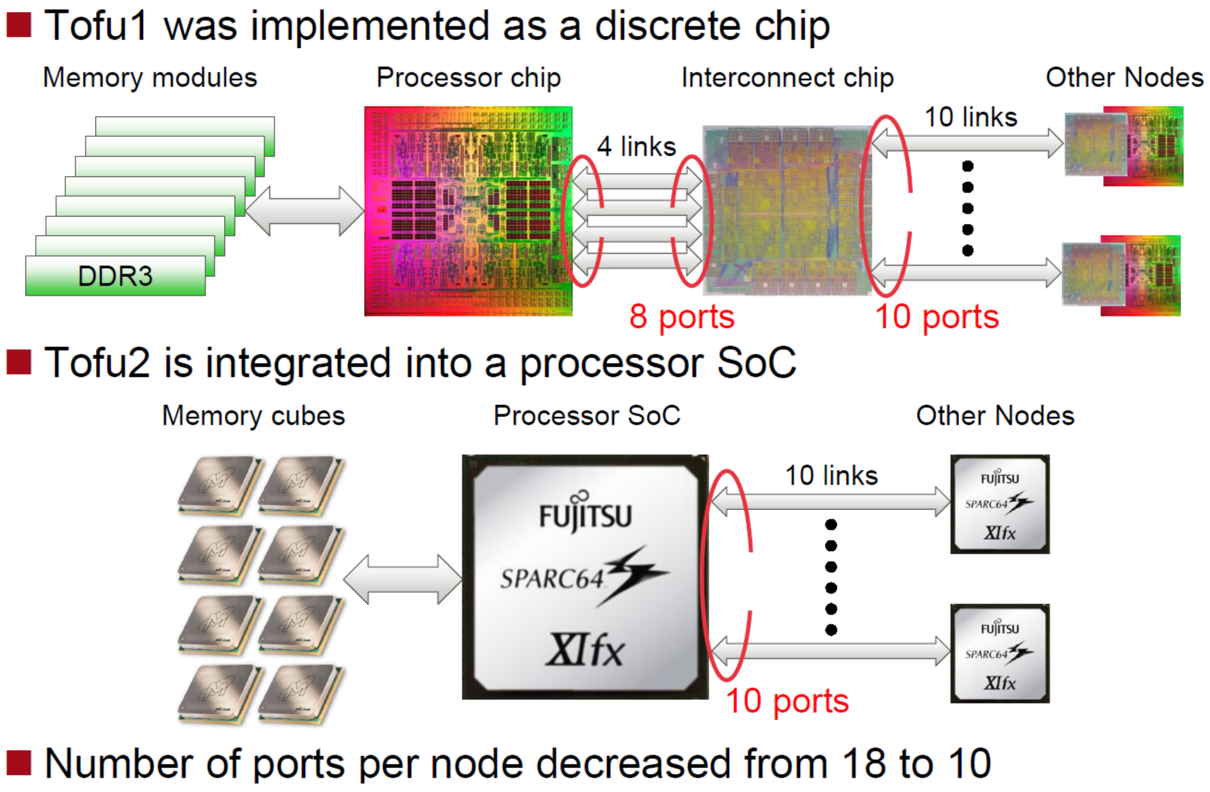

User Applications & Third Party Libraries
C and Socket libraries
Cray MPI
Shmem, UPC, Chapel, Coarray Fortran
User level Network Interface (uGNI)
DMAPP
IP stack
Lustre Client & Network Driver (LND)
Kernel Network Interface (kGNI)
OS Bypass
IP over Fabric (IPoGIF)
GNI Core
Hardware Abstraction Layer (HAL)
Gemini / Aries NIC

Figure 39: ARIES SW stack overview

3.4.8 *Fujitsu TOFU-2*

TOFU-2 is an upgrade of the original TOFU interconnect used in the K-Computer and should be a part of SoC (System on a Chip) [77] implementation in the next generation of Fujitsu's machines currently called Post-FX10. This means that the CPU will include the interconnect in a similar way to Intel's OPA (see Figure 43).



- Tofu1 was implemented as a discrete chip

- Tofu2 is integrated into a processor SoC

- Number of ports per node decreased from 18 to 10

Figure 40: TOFU-2 vs. TOFU-1 architecture

3.4.9 *IBM BlueGene/Q*

With 19 systems this is the second most represented custom interconnect amongst current supercomputers tightly coupled with the BlueGene/Q platform. There hasn’t been any update in last years of the key features and IBM will not further develop this interconnect. For upcoming systems like CORAL IBM will switch to Mellanox Infiniband EDR [78].

3.4.10 *TH Express-2*

Last but not least in size is the interconnect used in the currently biggest HPC system, the Chinese Tianhe-2. The NICs ASIC is connected via standard PCIe gen. 3, so we can expect adoption in other, even future, Chinese HPC systems (see Figure 41). It features modern characteristics like RDMA, offload for collective operations, address translation and memory protection and RAS features including CRC checks on link level.

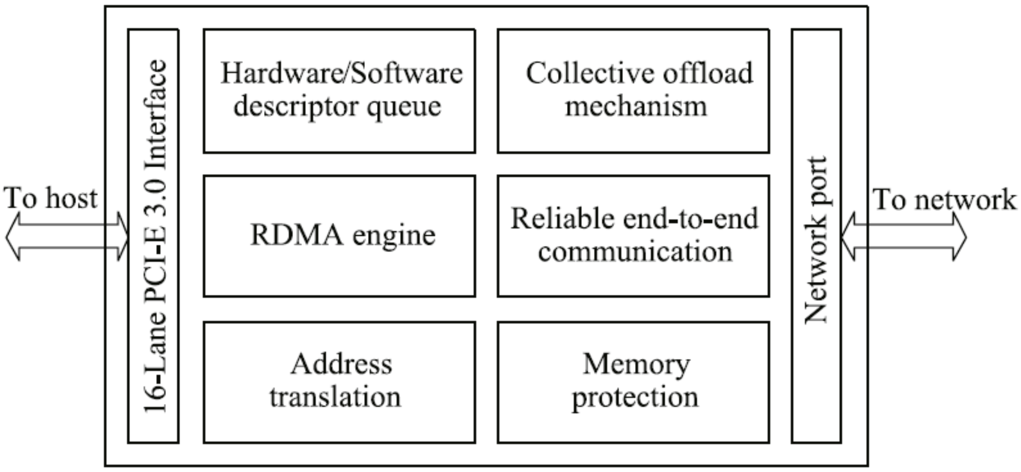

To host
16-Lane PCI-E 3.0 Interface
Hardware/Software descriptor queue
Collective offload mechanism
RDMA engine
Reliable end-to-end communication
Address translation
Memory protection
Network port
To network

Figure 41: TH Express-2 NIC block diagram

The software stack called GLEX2 (Galaxy Express-2) includes kernel drivers for Linux, user level library (glex), TCP/IP driver (gnet), PXE driver (gpxnet) and own MPI implementation called MPICH2-GLEX2 which is based on Argonne National Lab's MPICH2. The kernel modules support direct GPU-to-GPU (using NVIDA kernel patch for zero copy RDMA of CUDA pinned memory). There's also assisted support for direct NIC access from an Intel Xeon Phi coprocessor.

Minimum MPI latency is around 1.26 us and maximum peak bandwidth can be up to 20.6 GB/s for message sizes greater than 8 Kbytes using bi-directional RDMA PUT operations [79].

4 Solutions and architectures

After a brief overview of where main vendors are standing or heading for, this chapter devotes a section to some storage systems trends, then to cloud and virtualisation trends and experiences.

4.1 Vendors overview

4.1.1 *ATOS/Bull*

ATOS has now clearly finished the Bull acquisition that was announced in 2014. Bull is now the expert brand for Atos technologies, hardware and software products, which encompasses, but is not limited to, HPC and Big Data technologies [80].

Bullx series of HPC system has been continuously updated, since ‘first of a kind’ TERA100 system in 2010, with different flavours of nodes w/ and w/o accelerators (TESLA K40/K80 or Phi), SMP options of supernodes up to 288 cores and 24 TB RAM and service and I/O specialized nodes. Direct Liquid Cooling (DLC) is also an option, as well as ultra-capacitors embedded in the nodes (that can replace outer UPS units, as a protection against short current cuts, up to a few tenths of a second). It can be noted the bullion series of enterprise SMP servers benefitted from HPC developments and bullx R&D, and was another successful product line. In addition a number of references in Europe, it can be noted a recent Petascale contract for ATOS/Bull in Brazil [83].

In 2015 and more particularly during SC13, ATOS/Bull announced their new generation of HPC architectures, code name SEQUANA, described as the path towards Exascale. The approach encompasses strong co-design with CEA [81][82].

The fundamental brick is a “cell” comprising two compute cabinets, interconnect switches, redundant power supply units, redundant liquid cooling heat exchangers, distributed management and diskless support. Compute cabinets comprise 48 horizontal compute blades, with the associated power modules at the top of the cabinet and the redundant hydraulic modules for cooling at the bottom of the cabinet. 24 blades are mounted on the front side of the cabinet, while the 24 other blades are mounted on the rear side.

Each cell can therefore contain up to 96 compute blades, i.e. 288 compute nodes, equipped either with conventional processors (such as Intel Xeon processors) or accelerators (e.g. Intel Xeon Phi or NVIDIA GPUs). In each 1U blade, a cold plate with active liquid flow cools all hot components by direct contact – the Sequana compute blades contain no fan.

First blade options would be 3-node ones, each node with a 2-socket Intel Broadwell Xeon®, or 3-node, each node with an Intel Xeon Phi KNL.

Sequana is meant to accommodate different interconnect flavours (incl. Bull BXI cf section 3.4).

Figure 42: Bull Sequana cell

4.1.2 *Cray*

Cray continues to provide two product lines of HPC computing systems. Cray XC are optimized for very scalable high-end architectures, while Cray CS cluster computers provide higher flexibility. The currently largest systems based on Cray XC is the Trinity system at Los Alamos National Lab (US), with a peak performance of 11 PFlop/s, while the currently biggest Cray CS is a US-government system with more than 6 PFlop/s peak performance, which is currently listed at position #15 of the Top500 list.

The Cray XC (also known under the codename Cascade) was introduced in 2012. Cray's strategy is to provide a long lifetime for each platform. The next generation platform with codename Shasta has already been announced without details being disclosed. It is expected to become available in 2018 when the next Argonne Leadership Computing Facility (ALCF) is deployed.

Cray furthermore continues to provide products optimized for data analytics. The Urika-GD product line based on the custom Threadstorm4 Graph Accelerators was complemented by Urika-XA, an x86-based architecture announced in 2014.

4.1.3 *Fujitsu*

Fujitsu continues to pursue two HPC product lines. Better know is the high-end product line PRIMEHPC based on SPARC processors designed by Fujitsu. The K computer continues to be the flagship installation, which entered the Top500 in 2011 at position #1 and as of November 2015 was at position #4. The other product line, PRIMERGY, is based on x86 processors from Intel. Following the Top500 metric, the currently largest installation with a performance of about 1 PFlop/s is installed at Kyushu University.

Already in 2013, Fujitsu announced a next generation of PRIMEHPC architectures [89]. The planned enhancements included a new processor, SPARC64 VIIIfx, with wider SIMD units, use of the new high-bandwidth memory technology HMC (see chapter 3) and an updated version of the network, now called Tofu2. Prototype boards with 3 processors and 8 HMC memory stacks per processor have been show since a few years, but no large-scale installation based on this architecture has been announced, yet.

The x86 product line optimized for x86 runs under the name PRIMERGY CX [87]. 2U high chassis can currently host up to 4 half-wide dual-socket servers. These servers are available also with direct liquid cooling based on Asetek technologies [88]. This allows up to 80 servers with 160 processors to be integrated in a single rack. At SC16 Fujitsu showed at its booth single-socket servers comprising Intel's next generation Xeon Phi (Knights Landing).

Figure 43: Fujitsu density-optimized PRIMERGY CX chassis

4.1.4 *IBM*

IBM is currently fully re-organising its HPC product portfolio after it announced the end of the Blue Gene architecture line and after having sold the x86 business unit to Lenovo. Future IBM HPC offerings will be based on GPU-accelerated nodes. These will be the building blocks for the pre-exascale systems at Oak Ridge National Lab and Lawrence Livermore National Lab called Summit and Sierra, respectively, which will be deployed in 2017/18.

A first step towards a new HPC product line was taken by bringing a scale-out server based on POWER8 processors to market. The IBM Power System S822LC (codename Firestone) is a dual-socket system, which can host up to 2 NVIDIA GPUs. The next generation with codename Garrison is expected to become available during 2016. It is based on a modified version of the POWER8 processor, called POWER8', which features new NVLink ports, through which the processor can be tightly integrated with NVIDIA's next generation GPUs called Pascal.

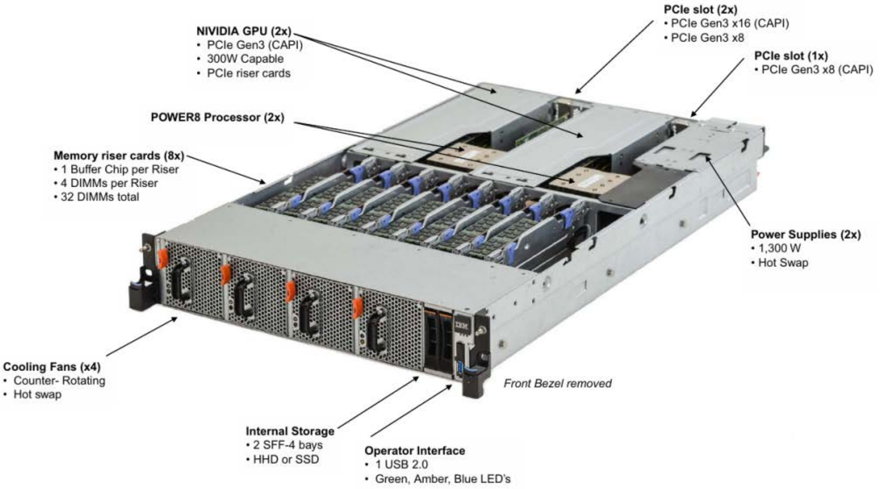


Figure 44: IBM Firestone server [90]

4.1.5 *NEC*

NEC has a long history with regard to supercomputing in Europe. While the company is also providing x86-based LINUX-systems to many European customers, the LX-series of scalar scalable systems, NEC's is also known for the product line of high-end vector machines. Its current implementation, the SX-ACE, is successfully used for example at HLRS, AWI or the CA-University in Kiel.

Vector computers have served the European scientists for many years, for example in Meteorology (UKMO, Meteo-France, DWD, DKRZ, CHMI) but also at universities on a European scale in France, the UK, Italy, Austria, Denmark, Germany etc. During the last decade the tendency for scalar processor based systems has limited the number of such computers in Europe. To enhance performance, the interest in accelerators and GPUs grew significantly, with non-trivial effort to port and optimize applications.

NEC's future products have the objective to use sustainable standards for programming and portable parallelization paradigms like MPI and OpenMP. They will rely on NEC compiler technology to answer to the expectations of scientific applications for code efficiency and sustainable standards.

The new systems will be tightly integrated into a scalable LINUX-environment and will implement a real vector-architecture with high memory bandwidth for a higher efficiency single computational level, while scalability will be a design-target of the overall system.

Availability to market target is early 2018. NEC announces a much more affordable price than previous NEC vector products and should therefore address a much wider market and be part of the European HPC-ecosystem.

4.1.6 *LENOVO*

Lenovo entered the HPC business by acquiring IBM's System x branch in October 2014 including the responsibility for 35 systems in the Top500 installed by IBM and based on Intel x86 CPUs – most prominent the SuperMUC (Phase 1 + 2) at the Leibnitz Rechenzentrum (LRZ) in Munich.

At the beginning most customers were sceptical about the direction Lenovo would take and especially US government agencies had security concerns with Lenovo as a Chinese company. To resolve those doubts Lenovo went successfully through an intensive US government security audit certifying the whole supply chain to be safe and back-door free.

Lenovo's latest product announcements and roadmap show a clear strategic interest in the data centre market with modular hardware also suitable for HPC systems. Lenovo offers hardware based on Intel's high-end product portfolio (Xeon, Xeon Phi, Omnipath, 3D-Xpoint) in classical air or (high temperature) direct water cooled systems.

Lenovo also committed its interest in HPC by actively participating in the SPXXL user group of (former IBM) HPC centres and is going to install a large 15 PFlops system at CINECA in 2016/2017.

4.1.7 *HPE*

Hewlett Packard decided in 2014 to split the company into two separate companies, Hewlett Packard Enterprise (HPE) and HP Inc. (HP). This change was effective as of 2015-11-01.

All systems relevant to the HPC market were moved over to HPE, so for the typical HPC customer the main difference so far is the replacement of the name HP with HPE. Judging by what was presented at HP-CAST and SC'15 in Austin no major changes are due for the HPC market either.

HPE continues the strategy of providing a broad spectrum of different systems, both smaller ones and those aiming for the upper part of the TOP500 list. However, the current systems are not targeted at the very highest end of the TOP500. Only six of the 155 HPE systems on the TOP500 are among the top 100 systems, with the highest-ranking system at position 38.

Apart from standard rack servers HPE also has high-density systems specifically aimed at the HPC and data analysis market. Earlier this was the BL and SL series, but the BL servers are now more marketed for virtualization with the Apollo line replacing the SL systems. Given the popularity of the blade servers the BL servers will probably still be used for HPC systems, especially for commercial systems. With the exception of the Apollo 8000 discussed below, the rest of the Apollo systems might be seen as continuations of older product series or packaging them for HPC. In general the Apollo servers become more blade like with increasing model numbers, Apollo 6000 integrates the management in the chassis and Apollo 8000 also has integrated IB for example.

In recent years HPE has been moving gradually towards the higher end with the introduction of the water-cooled Apollo 8000 range; the current TOP500 has one such system (at #39). If HPE will make inroads into the market for really large systems, it will probably be the Apollo 8000 (or any successor) that is the likeliest system to be installed. A potential challenge for HPE in the future will be to maintain the broad presence among the more mass-market HPC customers, while also aiming for the high end systems.

4.1.8 *SGI*

References for this subsection: [84][85][86]

SGI UV

The SGI UV symmetric multiprocessing (SMP) systems provide cache-coherent shared memory for the most compute- and data-intensive application workloads. These Linux- and

Intel-based servers utilize SGI NUMAlink interconnect technology. The UV systems may be leveraged as a "super nodes" for clustered HPC.

The SGI UV 3000 scales up to 256 CPU sockets and 64TB of cache-coherent shared memory in a single system. Enabling such capability is 6th generation SGI NUMAlink ASIC technology interconnect, the MPI Offload Engine (MOE) is meant to optimize applications developed for distributed computing - similar in concept to a TCP/IP Offload Engine (TOE), which offloads TCP/IP protocol processing from system CPUs. Designed for smaller, compute-intensive environments, SGI UV 30 is a 2U, 4-socket server providing up to 3TB of in-memory computing power.

SGI ICE

Compute nodes feature the dual socket Intel Xeon processor E5-2600 v3 series. Nodes can be further augmented with Intel Xeon Phi Coprocessors or NVIDIA GPU Accelerators. Support for Xeon Broadwell and Skylake family processors is planned, as well as for single socket standalone KNC platform. Alternate processors vendors might include ARM and AMD companies. SGI ICE XA can provide up to 191 teraflops per rack and grow to tens of thousands of nodes. Quad node or dual node blade types are available. Blade enclosures provide power, cooling, system control, and network fabric for up to 9 compute blades via an integrated midplane, and up to four enclosures in a single rack. SGI ICE XA utilizes industry-standard InfiniBand networking with complete flexibility in topology. The ICE XA line may be extended to support Intel Omnipath in future. Standard Linux distributions can be provisioned. SGI ICE Systems are DLC cooled.

SGI InfiniteData cluster

SGI InfiniteData Cluster combines high server and storage density with scale-out performance, designed for Big Data Analytics and moderate technical computing workloads.

InfiniteData Cluster offers up to 1,920 cores and 1.9 PB of data capacity per rack. The systems are air or air-rear door heat exchanger cooled. All InfiniteData Cluster components-trays, interconnects, and I/O-are cold-aisle accessible to support, hot-aisle/cold-aisle floor plans.

SGI Rackable

SGI Rackable are standard-depth, rackmount dual-socket servers, with either Intel or AMD processor architectures (in particular E5-2600 v2 and v3 family processors and the AMD Opteron processors) and support up to 1.5TB of memory per node in 24 slots. SGI Rackable comes in 1U and 2U form factors. The systems are air or air-rear door heat exchanger cooled.

4.1.9 *D-WAVE*

The company

The company D-Wave was founded in 1999 and employs 100 people. They have registered more than 100 patents. D-Wave is the only vendor of computing machines based on quantum technology, while the other companies focusing on quantum are still in the research project.

Four systems are installed, one shared by Lockheed and University of Southern California, one shared by Google and NASA Ames (for which a long term agreement was signed: the machine will follow the developments in D-Wave Technology), one in a classified area in the US and the last order was issued in November by Los Alamos National Laboratory.

The technology

D-Wave technology uses quantum effects of supra-conductive loops and the last processor D-Wave 2x relies on 1024 "qubits".

The operating environment is extreme. The quantum processor requires: high vacuum, a temperature close to absolute zero (0.02 K), high isolation of the magnetic field and vibration. The energy consumption of around 15KW, is mainly (we may say entirely as supraconductor consumption is none!) dedicated to cooling.

Scope of the D-WAVE Technology

The instruction set does not include arithmetic operations on numbers. Optimization problems, graph courses, minimization of energy systems research can be solved with accelerations of up to 5 orders of magnitude. However, the programming model is still the subject of research and development and it is very difficult to build a model for a problem. Nevertheless, Google claims that it enables tremendous acceleration for image recognition. Some of the domains targeted by D-wave technology: Monte-Carlo simulation, machine learning, compression, object detection, optimisation problems

4.2 Storage Organization for Exascale Computing Systems

Storage trends

General trend in the direction of software becoming more important on the storage side, with for example de-clustered RAID decreasing the importance of traditional hardware RAID controllers. Vendors place more emphasis on integrated solutions with a software stack included to provide more of turnkey systems rather than hardware building blocks.

While much of the focus is on the high performance storage options, rotating hard disks are still the cheapest option for volume storage with 6 and 8 TB drives becoming common in storage systems. Shingled drives (SMR) have not yet started to appear on a large scale in products.

Lustre and GPFS remain the most common clustered file systems in use at HPC sites. Multiple other file systems are being used, but mainly Lustre has a large market share. File systems coming from a media (video etc.) background are trying to enter the HPC market, just like GPFS once did.

Node local storage seems set to remain for the medium term future, with persistent memory being used only for high end compute nodes. With interconnects and file systems available, combining local storage from multiple nodes into transient job wide storage is becoming a viable option.

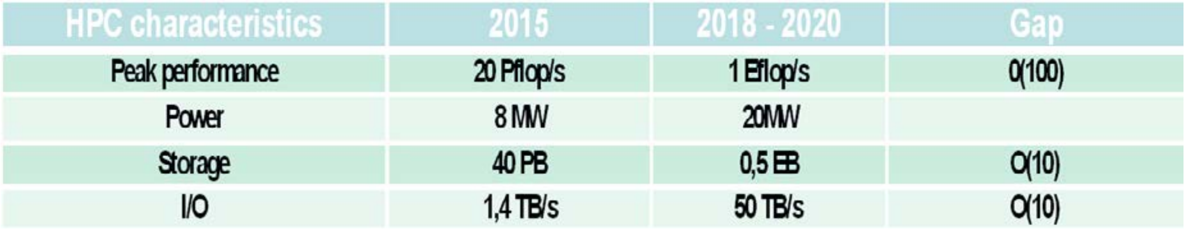
Lustre governance

In recent years the major Lustre vendors, with the exception of Seagate[1], have started using the Intel Enterprise Edition for Lustre distribution as the base of their products. Some Lustre users have felt that the development of Lustre is currently done less in the open than previously, and an impromptu “emergency BoF” was called at SC’15 by Stephen Simms from Indiana University to discuss this. There seems to be a wide range of opinions on what the purpose of OpenSFS should be, and how the composition of the board should be. Historically OpenSFS played a larger role as a way to pool resources together for Lustre development, but this is now handled by the vendors selling Lustre storage systems. Some see this as affecting

[1] On April 5th, 2016, date of (almost) final writing of this report, Seagate announced that they will also be using the Intel Enterprise Edition for Lustre in the future. See http://www.seagate.com/gb/en/about-seagate/news/seagate-offers-powerful-new-lustre-choice-for-hpc-storage-master-pr/

the stability of the open source Lustre repository with the master branch being the focus, and that the formal open source releases are not stabilized.

Exascale challenges for I/O



HPC characteristics	2015	2018 - 2020	Gap
Peak performance	20 Pflop/s	1 Eflop/s	0(100)
Power	8 MW	20MW	
Storage	40 PB	0,5 EB	O(10)
I/O	1,4 TB/s	50 TB/s	O(10)

Figure 45: exascale I/O challenges

Multi-petascale and exascale platforms will have to manage the discrepancy between computing and I/O capability. The gap may be managed by either tiering approaches or new designs in HPC solution to reduce the data movement. Examples are shown in next two sections

4.2.1 *Reducing data movement*

Example of SGI Zero Copy Architecture

Data growth in High Performance Computing (HPC) will continue to Exascale levels and beyond. Accordingly, it’s becoming too costly to have copies of that data and, too energy intensive to move them. Currently, many workflows are as follows:

$$Compute\ > Store\ > \boldsymbol{Copy}\ > Analyze$$

First, the computational result delivered by HPC is stored in the external storage over a cluster communication fabric, which requires copying the data to the memory of the storage server (burst buffer) before it can be forwarded to the actual storage device. This creates additional storage load on the cluster communication fabric, reducing the effective utilization of computational resources.

Further, often the computing resources are very different and/or separate from the analysis resources. External storage becomes the default common meeting place.

The analysis portion begins by recalling the data back from the external storage into the memory or local storage of High Performance Data Analysis (HPDA) system.

The novel concept of SGI Zero Copy Architecture (ZCA) has been identified in order to remove the costs induced by the latencies of the burst buffer and stores/loads to the external storage. The ZCA inserts a non-volatile memory (NVM) device in place of burst buffer and storage, see Figure 46. The external storage is no longer the primary location where active data are handled; its role is offset to long term/archiving purposes.

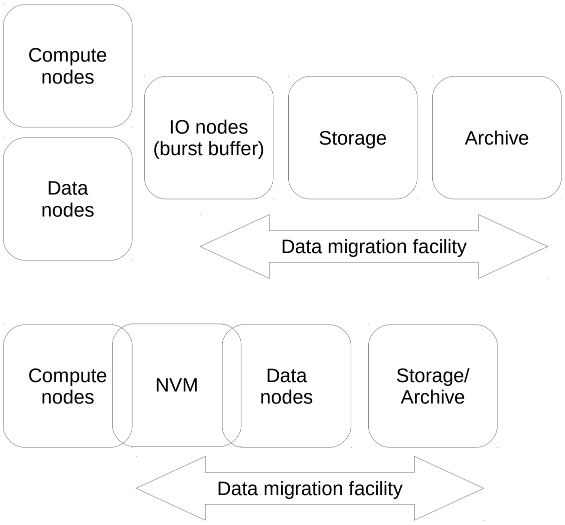


Figure 46: SGI Zero Copy Architecture

Upper half: Traditional HPC/HPDA architecture featuring burst buffer, storage and archive.
Lower half: SGI ZCA. The NVM device is local to both Compute nodes (HPC) and Data nodes (HPDA), offsetting the burst buffer and storage.

Deploying the SGI ZCA, the cluster fabric is no longer used by the Compute nodes to access storage; it exclusively carries the HPC data exchange.

The NVM device is shared among multiple Compute nodes and Data nodes, connected as local storage to both, via the local data fabric (PCI express). The NVM device itself contains a number (M) NVM modules and a local data fabric switch (PCI express Lane Switch). For scalability, multiple Data nodes are interconnected by a non-local (NL) storage fabric, thus bridging multiple NVM devices, see Figure 47. The bridging communication within the Data nodes may be constructed as a zero copy access with no CPU or local memory involvement. The Data nodes are further attached to an External storage and Archive facility for longer term data storage purposes.

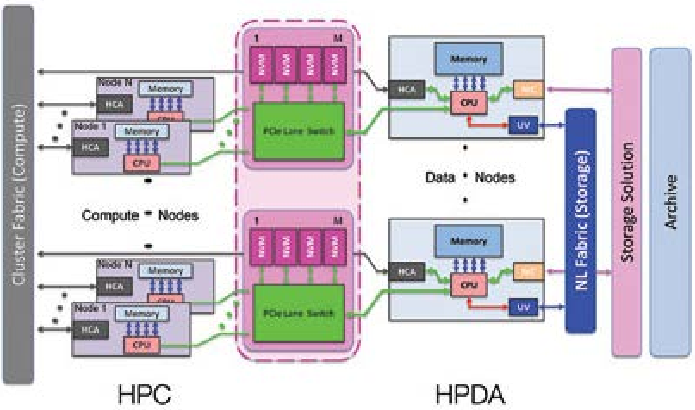


Figure 47: SGI Zero Copy Architecture

The NVM devices act as a local storage to both HPC and HPDA.

Each Compute node writes locally for performance, yet can access others' data globally, see Figure 48.

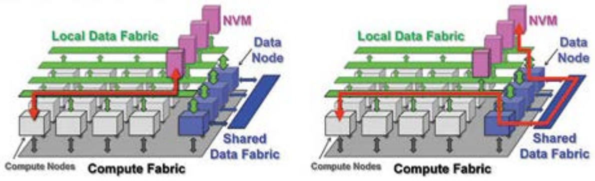


Figure 48: Data appear local, even if it's not

The result is the ability to perform typical workflows, in situ analytics, visualization and other processing at the highest performance and lowest latency without creating interim in-memory data copies or involving I/O acceleration techniques such as burst buffers.

SGI ZCA can be used to create a true "no copy" workflow:

$$\mathit{Compute}\ > \mathit{Store\ Local}\ > \mathit{Analyze}$$

The computing resources write to storage as they typically would, but that storage is now local/internal NVMe devices. Compute processes simply request direct I/O to transfer data between process memory and storage via single direct memory access (DMA) transfer. The next step/process in the workflow can also see that storage as local and proceed as needed, without retrieving a copy over the network.

4.2.2 *Tiering Solution*

The introduction of SSD in HPC landscape reduces the gap between memory and traditional non-volatile storage solutions.

From SAGE (Percipient StorAGe for Exascale Data Centric Computing) H2020 research project [143], a tiering layout in Figure 49 below.

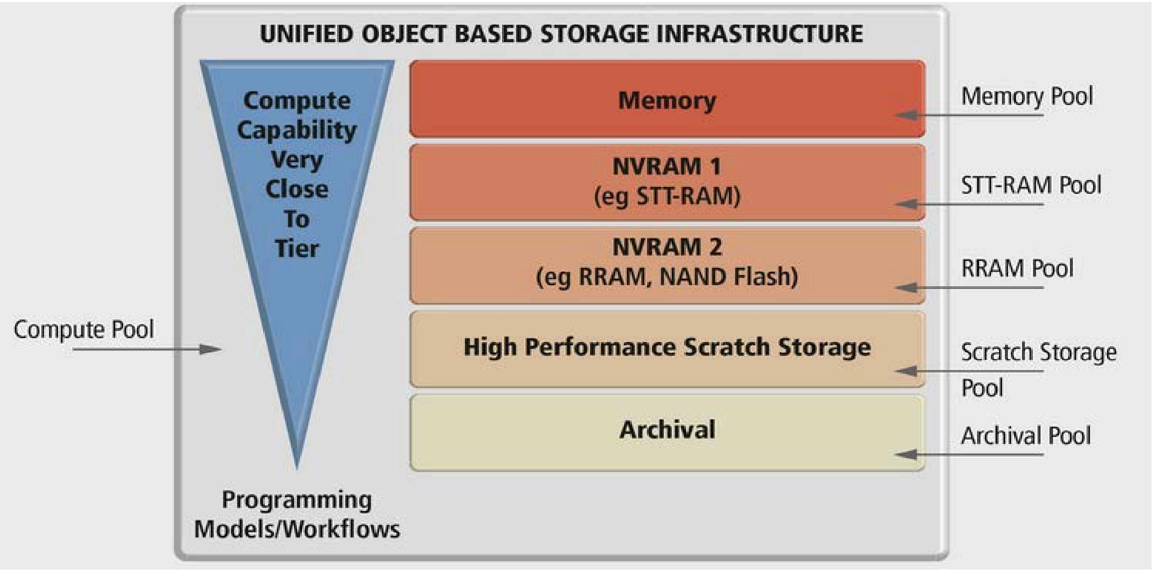
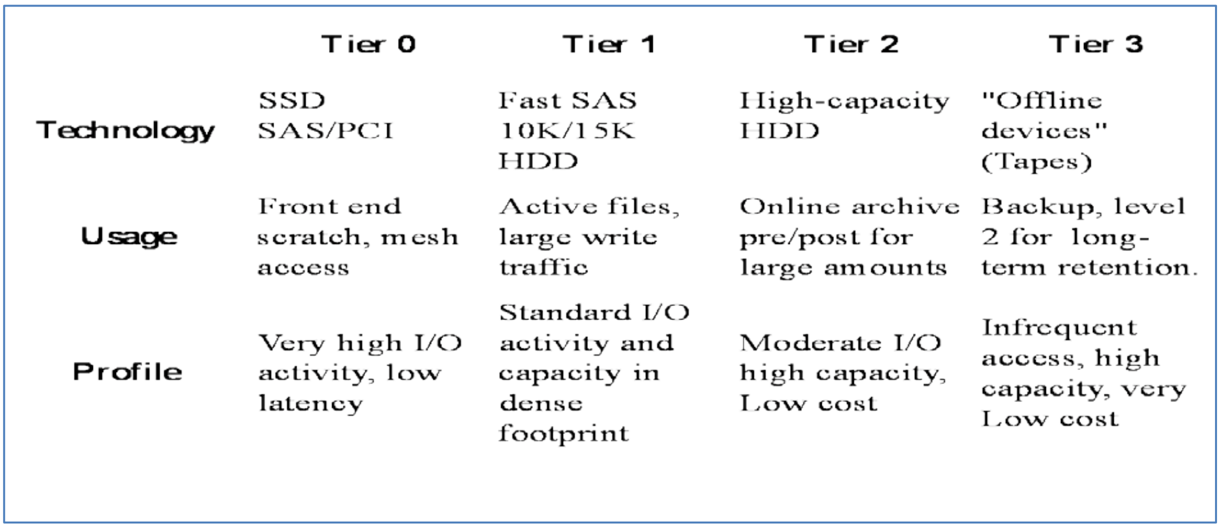

UNIFIED OBJECT BASED STORAGE INFRASTRUCTURE
Compute Capability Very Close To Tier
Memory
NVRAM 1 (eg STT-RAM)
NVRAM 2 (eg RRAM, NAND Flash)
High Performance Scratch Storage
Archival
Memory Pool
STT-RAM Pool
RRAM Pool
Scratch Storage Pool
Archival Pool
Compute Pool
Programming Models/Workflows

Figure 49: SAGE tiering layout

Tiering usage in HPC facilities



	Tier 0	Tier 1	Tier 2	Tier 3
Technology	SSD SAS/PCI	Fast SAS 10K/15K HDD	High-capacity HDD	"Offline devices" (Tapes)
Usage	Front end scratch, mesh access	Active files, large write traffic	Online archive pre/post for large amounts	Backup, level 2 for long-term retention.
Profile	Very high I/O activity, low latency	Standard I/O activity and capacity in dense footprint	Moderate I/O high capacity, Low cost	Infrequent access, high capacity, very Low cost

Figure 50: Tiering usage in HPC facilities

SSD integration examples: by DDN, Infinite Memory Engine

IME enables the infrastructure to separate the provisioning of peak from sustained performance requirement, unleashing the performance of the parallel filesystem from the number of disks drives which, most of the time, leads to a significant amount of unused storage – and delivers a better performance/cost answer. (See "IME typical peak provisioning below")

IME is between compute nodes and PFS, while I/O driver "IME aware" is installed, intercepting application I/O for rearranging: stripping, alignment. (See DDN Infinite Memory Engine below in Figure 51).

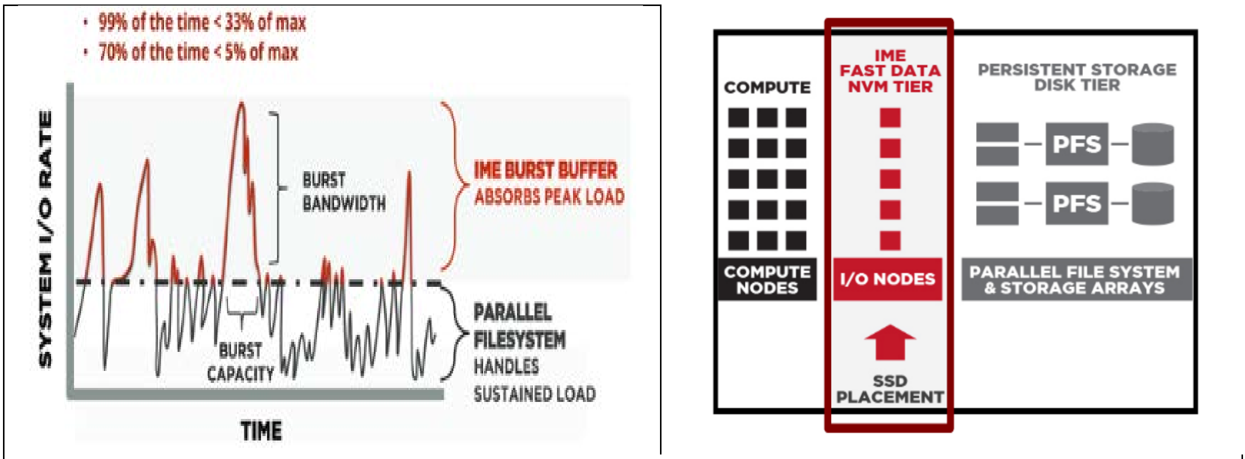
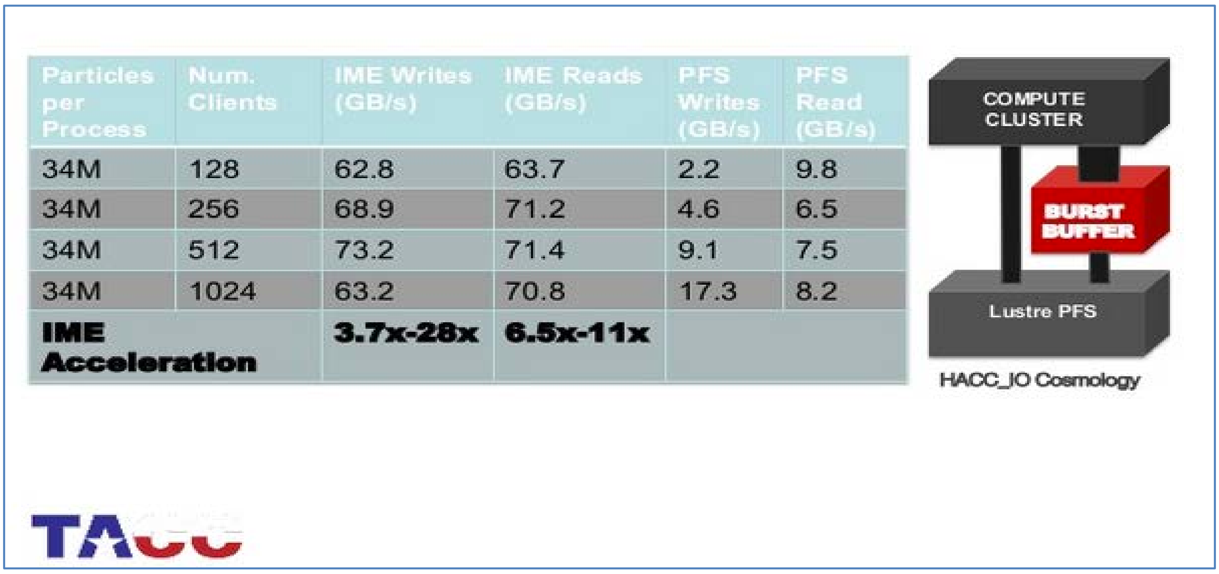

• 99% of the time < 33% of max
• 70% of the time < 5% of max
SYSTEM I/O RATE
BURST BANDWIDTH
IME BURST BUFFER ABSORBS PEAK LOAD
BURST CAPACITY
PARALLEL FILESYSTEM HANDLES SUSTAINED LOAD
TIME
COMPUTE
IME FAST DATA NVM TIER
PERSISTENT STORAGE DISK TIER
PFS
PFS
COMPUTE NODES
I/O NODES
PARALLEL FILE SYSTEM & STORAGE ARRAYS
SSD PLACEMENT

Figure 51: left - IME typical peak provisioning / right - DDN Infinite Memory Engine

IME acts as a smart buffer, improves small I/O, improves read and write by realigning.
MPI – I/O support is implemented and an API is available. Figure 52 shows an application result at TACC [144].



Particles per Process	Num. Clients	IME Writes (GB/s)	IME Reads (GB/s)	PFS Writes (GB/s)	PFS Read (GB/s)
34M	128	62.8	63.7	2.2	9.8
34M	256	68.9	71.2	4.6	6.5
34M	512	73.2	71.4	9.1	7.5
34M	1024	63.2	70.8	17.3	8.2
IME Acceleration		**3.7x-28x**	**6.5x-11x**		

COMPUTE CLUSTER
BURST BUFFER
Lustre PFS
HACC_IO Cosmology

Figure 52: Acceleration by IME – Cosmology application at TACC

SSD integration examples: by Seagate, CL300 Small block accelerator

Seagate has chosen an implicit implementation of SSD acceleration. The I/O driver on the host is not aware of the underlying acceleration implemented by SSD tiering - see **Figure 53**.

Each Storage Unit embeds SSD devices

Seagate « Nytro XD » cache management software
:

- Monitors writes Block Stripe Size
  Small Blocks Write to SSDs
  Large Blocks Write to HDDs
- Up to x16 on mixed I/O and small file performance
- 16 GB/s random read/write by SSU

Figure 53: Seagate CL300 Small block accelerator

Summary of tiering and trends at both DDN & Seagate :

- DDN: IME provides a wide range of improvements but implies a software dependency and scales independently from backend storage;
- Seagate: CL300 Small block accelerator – Seagate focuses on block size but requires not host software dependency;
- DDN SSD acceleration at backend level is part of 2016 roadmap;
- Seagate: “Burst Buffer” independent unit is part of 2016 roadmap;
  Both DDN & Seagate will include implicit and explicit tiering acceleration;
- DDN & Seagate: Fine grain acceleration i.e. SLA at file level by tagging is on study.

4.3 Trends in cooling systems

One trend in high-performance computing, for many years, has been for increased power density and that trend is expected to continue for some time. To be able to cool the electronics in high-density computing system most vendors today provide solutions for liquid cooling. It can be of different forms and with different characteristics. Another trend in cooling is the goal to provide an as high temperature of the outgoing liquid as possible. This makes it possible to use free cooling most of the year also in warmer climates and opens the possibility for heat re-use in colder climates. To achieve high temperatures, it is important to catch the heat close to the source before it becomes too diluted. Also, the number of steps in the transfer (for example heat exchangers) should be minimized to avoid loss of temperature.

A simple form of liquid cooling is hybrid cooling where the electronic components are cooled by air and the air is cooled by water quite close to the source. The water-cooling of the air is normally done by heat exchangers in the same rack as the electronics for example as cooled doors or as a more special design. Hybrid cooling has the advantage of being simple to design and the electronics can be the same as in air-cooled versions. However, some temperature is lost when converting from air to liquid and the density is somewhat limited. But due to the simplicity, we expect this to be used frequently also in the future.

Direct liquid cooling where the electronic components are cooled by the liquid directly is becoming more and more used. One reason is the higher densities mentioned above and the high outgoing temperature that allows for free cooling in most climates and heat re-use in colder climates. Many vendors provide direct liquid cooling today with more to come. The drawback is slightly higher investment cost for the supercomputer. It may be compensated by lower costs of the data centre infrastructure as it may be simplified. Operational costs can also be lower than air-cooling, especially in countries with a warm climate. In many cases also, direct liquid cooled systems require additional air-cooling if not all the components in a system are water-cooled.

Submerged cooling has been around for some time without reaching too much momentum. Submerged cooling is when you take one or several boards of a system and submerge them in some liquid other than water in a closed or open container. Submerged cooling has the advantage of cooling all the components on the submerged parts and the resulting output temperature can be excellent. In some cases, the boards and the containers are special made and that adds costs to the system. In other cases, you take ordinary boards and place them in an open container, which is of course more cost efficient. Maintenance can, however, become more difficult and the long-term effects on standard components exposed to the fluid used are unclear. Operational costs and possibly economic benefits are similar to the direct liquid cooling case.

Another trend in the cooling of high-performance systems is the use of other liquids than water in hybrid cooling and direct liquid cooling. In some cases, the liquid can undergo a phase change (boil) and this can take up a large amount of heat. Temperature loss can be very small. Also, using non-conductive fluids can lower the risk of having leakages in the system. So far there are no long-term tests of large-scale deployments that let us draw conclusions about this kind of solution. Some of the fluids used can have a negative environmental impact during manufacturing and destruction. In addition, the usage of non-conductive liquids that may change phase at low temperatures (30-80 °C) is posing additional technical challenges for the equipment. Physical properties of this kind of liquids differ significantly from normally used coolants; therefore, active parts (e.g. pumps) of the cooling loop may wear out quickly or behave in a manner that is not tested by the manufacturers.

4.4 Trends and return of experience in virtualisation and cloud delivery

4.4.1 *Cloud*

Cloud HPC is becoming an increasingly standard offering in the product portfolio of HPC centers. The ability to self-provision virtual clusters provides an alternative for bare-metal hosting and flexibility for customers to implement customized software stacks.

Below we provide two case examples of modern cloud architectures at the work package partner sites, CINECA and CSC.

CINECA PicoCloud

The HPC cloud infrastructure currently installed at CINECA is based on Openstack [91] Kilo, the 11th release of the open source software for building public, private, and hybrid IaaS clouds.

The Cineca Openstack infrastructure consists of a physical server, which host the Cloud Controller, the central management system for OpenStack deployments, and the other 15 physical servers with the role of Compute Nodes, providing the processing, memory, network and storage resources to run virtual machines.

The 16 servers that constitute the physical infrastructure are all equipped with two Intel Xeon E5-2670v2 2.50 GHz, 10 cores each, 128 GB RAM, a HCA Mellanox Connect-X3 FDR and they have RDMA access to a shared storage area based on GPFS.

Apart from an Ethernet interface dedicated only to the server management, the 16 servers do not have additional physical Ethernet interfaces. Therefore each server is also equipped with the Mellanox driver eth_IPoIB, which provides a standard Ethernet interface over IPoIB to be used as a Ethernet Physical Interface (PIF). This PIF can serve one or more Virtual Interfaces (VIF) and the eth_IPoIB driver supports L2 Switching (Direct Bridging) as well as other L3 Switching modes (e.g. NAT). In this way the servers are fully configured to host an external network and a data network mutually independent, as required to deploy OpenStack with tenants isolation.

The OpenStack deployment on the physical servers has been carried out using RDO [92] on CentOS Linux 7.2, through the PackStack [93] installer, configured to install Nova, Glance, Cinder, Neutron, Horizon, Heat and Keystone2 on the Cloud Controller and all the Compute Node with KVM hypervisor.

The network services has been configured to use Neutron Distributed Virtual Router (DVR) [94], in order to avoid that all L3 traffic was sent through a single network node, including even traffic between VMs residing on the same physical host, as well as a single point of failure (SPOF) of the architecture.

[2] Openstack *Nova* mainly manages computational resources; *Glance* manages and provides VMs images; *Cinder* handles block storage devices that can be attached to VM instances; *Neutron* manages network resources; *Horizon* is the Openstack's dashboard and provides a web based user interface to OpenStack services; *Heat* implements an orchestration engine to launch multiple composite cloud applications based on templates and *Keystone* is a service shared with all the other mentioned services that provides authentication and authorization functionalities

The storage area for Nova, Glance and Cinder services are configured on a GPFS fileset shared among all the Compute Node, to minimize the VMs migration time between servers and to ensure high performance access to the block storage resources to the VMs.

A first assessment of the computational performance3 of VMs showed that, for single core jobs, the performance inside a VM are substantially the same, and, for parallel jobs, we observe a decrease between 10 and 15% compared to bare-metal case; in the storage area satisfactory performance are observed, even if substantially lower than physical access, a result which is not surprising given the non-use of protocols such as iSER and SRP.

CSC cPouta and ePouta

The cloud offerings at CSC consist of two different services, ePouta and cPouta (Pouta means a mostly sunny sky with some Cumulus –type clouds). Both are based on OpenStack and as with CINECA are based on OpenStack and RDO.

cPouta

cPouta (Community Pouta) is a general-purpose community cloud that uses the compute nodes of CSC's Taito [95] cluster to run the VM instances. It has been operating in production use since mid-2014. During the operation the partition of nodes dedicated to cloud use has been growing steadily and now comprises of 1650 cores (103 nodes with dual E5-2670 2,6 GHz 8 core CPU and 256GB RAM).

A main difference between the cPouta and Taito nodes is that cPouta uses 40 Gigabit Ethernet instead of FDR InfiniBand. This design decision was made early in beta testing as it was found that lack of redundant operation mode in the InfiniBand to Ethernet gateways (BridgeX) caused a single point of failure. Actual hardware changes that this caused were limited to cable rerouting as the IB switches and network cards support Ethernet (special license required). The tight integration with the production cluster has proven useful as capacity can be added based on demand.

Currently cPouta only provides non-oversubscribed instances. This means each instance is guaranteed at least one dedicated CPU core. However, in practice many of the users actually run non-CPU intensive tasks such as web and database servers and having a dedicated core wastes resources. This is being addressed by the addition of oversubscribed instances in the near future.

The need for extra overhead and capacity management is also more critical as the concept of queuing and queue management does not exist in the same way as with HPC jobs. There always needs to be some extra capacity to ensure that users can launch virtual machines. For the very large flavors that occupy a complete node, this means that there should be full nodes available.

[3] This evaluation was carried out by running the *HPL* to measure the performance of the CPU and *Stream* and *IOzone* to measure the bandwidth towards memory and storage respectively.

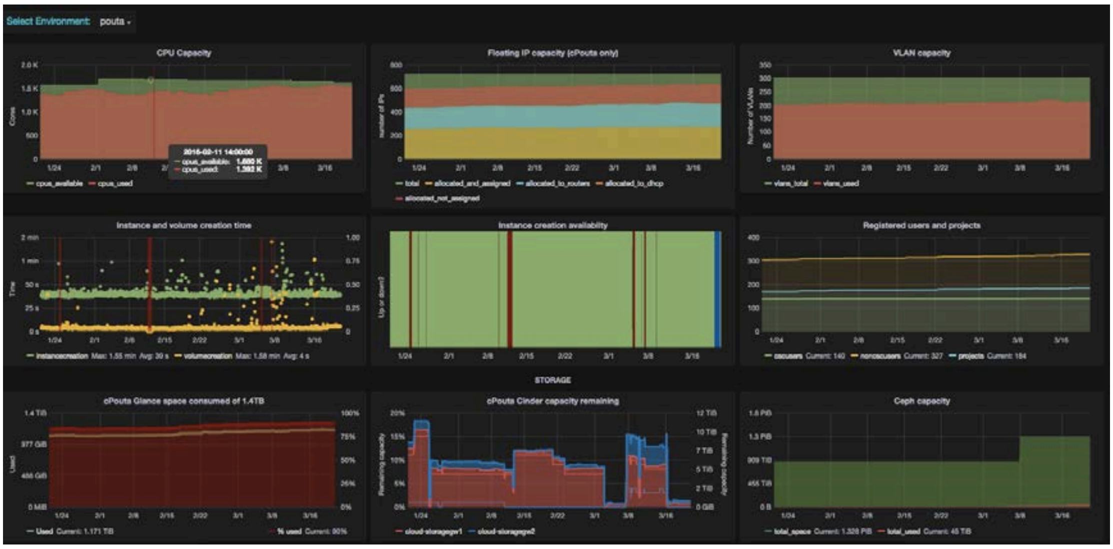


Figure 54: A Grafana dashboard to monitor cPouta capacity and performance

In addition to simply providing a new way of deploying virtual cluster instances, the cloud has proven extremely useful for rapid prototyping and agile development of services for researchers.

For storage, using volumes from CSC's DDN SFA12k enterprise storage system was originally used. At the time of writing this is being migrated to Ceph. The advantage is lower cost and a much more linearly scaling cost model. It will be possible to expand the storage in the same way as compute capacity whereas with large NAS servers, one needs to buy very expensive controllers as capacity demand increases. The complexity is reduced as there are less hardware, firmware and software layers involved. It is also possible to better accommodate individual users and user groups with specific SLA requirements for their storage (i.e. more replicas for better resiliency).

Being able to access the Lustre HPC storage and the batch job queuing system in a high-performance, reliable, secure and automated way from tenants is a key future development challenge. This would enable workflows where the cloud could seamlessly be integrated with the traditional HPC environment without the need to move data between different storage silos. A very long term strategy could be to have the HPC storage migrated to Ceph but it is still unclear when if ever Ceph will reach the performance in HPC workloads and MPI I/O that's comparable to Lustre.

Building platform services (PaaS) on top of the cPouta is in progress and some early example is the Pouta Blueprints service[4] that is used to launch applications on-demand with a simple web interface. There are also "plug-and-play" instructions in the form of configuration management files to deploy virtual clusters.

ePouta

ePouta ("Enterprise Pouta") is intended to provide virtualized datacenter expansion to institutional customers. Unlike cPouta, which is accessible from the general Internet, ePouta is made a seamless part of the customer network either via a Virtual Private Network (VPN) or an Optical Private Network (OPN, lightpath). This allows ePouta to be used for processing of sensitive information, such as genome data. Even in the case of a major server and/or firewall misconfiguration, the data is not visible directly to the Internet.

[4] pb.csc.fi

Experiences

Maintaining and operating a production-quality custom-designed OpenStack environment requires effort and a dedicated team. The cloud services at CSC are maintained by a cross-organizational team consisting of ~6 FTE working in a Scrum workflow. However, having this expertise in-house is beneficial also to use cases beyond HPC: Non-HPC customers as well as in-house developers are adopting cloud-native DevOps tools and practices for which the traditional virtualization and bare metal capacity platforms are not optimal. Ideally such a cloud development initiative in a large research IT services organization should be viewed in a larger company-wide strategic context and not just limited to HPC.

Allowing customers to set up and maintain VMs increases the risk of misuse and potential attack surface of the organization. Having very clear terms of use and security instruction documents as well as restrictive default settings on firewalls and OS images are good ways to mitigate these risks. Additional security services such as network vulnerability scans are also useful.

Customers need to also be educated about cloud best practices. Many come from backgrounds where jobs are either run with a batch job queuing system or static servers set up manually. Providing usage examples and webinars on modern cloud administration practices, such as configuration management (Puppet, Ansible etc.) as well as how to use the OpenStack CLI and API, improves the usage efficiency. Furthermore these skills are becoming a new norm in IT services and should be encouraged.

4.4.2 *Containers*

In the last few years, container technologies have been rapidly adopted in the IT industry and the technology has all but become synonymous with Docker [96]. Docker has taken the fairly mature concept of containers (Solaris Zones, IBM LPAR, LXC…) and rapidly developed it into a user-friendly product with a workflow that seamlessly fits the DevOps philosophy and modern microservices architectures.

Containers offer a restricted namespace for the application while sharing the underlying kernel infrastructure with the host.

The basic value proposition of containers is that they help to manage and run applications with complex dependencies easily and efficiently. This can be utilised in the context of HPC in a variety of ways. These cases assume a model where we are running the containers as tasks under a HPC batch job queuing system such as SLURM.

Currently containerised HPC is in its infancy but momentum is clearly building as evidenced by the following developments during 2015:

- IBM announced in spring that their LSF suite supports Docker [97]
- There was a well-attended workshop at ISC15 in July around the subject [98]
- More recently Cray announced nearly imminent support for Docker in SC15 in November [99]. This seems to largely leverage on the work that NERSC has done with their “Shifter” User Defined Images project [100].
- The recently published draft of the APEX2020 RFP for the next-generation Department of Energy systems has support for containers as a desirable feature [101]. This significant contract will likely motivate many vendors to accelerate the development of these features.

Some potential use cases for use of containers in HPC:

- **Root access needed:** Some application stacks assume that the user installing software has administrator level access on the system. Often this is infeasible at multiuser HPC sites due to security concerns. This can be typically circumvented, for example by editing installation scripts, but it can be difficult and error-prone. With containers, these applications can be installed in a sandbox with administrator privileges without any manual workarounds.
- **Complex environments:** Some application stacks can be very intricate and have a variety of dependencies to different uncommon libraries and/or have built systems which are complicated. Containers enable packaging these stacks with all the dependencies rolled together.
- **Non-standard Linux distribution requirements:** Some applications are dependent on a specific Linux distribution.
- **"Bring your own application stack":** One of the advantages of Docker is the simplicity one can share containers through the public DockerHub or a private repository. This enables users to deploy and run containers of their choice on different systems with ease using the simple command-line tools.
- **Preserving the stack for reproducibility:** Scientific papers often provide insufficient information about application runs to ensure reproducibility of computational results. Typically the main system name and software versions are included but rarely any detailed information on intermediate libraries (i.e. MPI, numerical libraries) and their versions. This can make reproducing the results extremely difficult, especially as HPC centers may clean up and retire old versions of libraries with little notification. Reinstalling these may be difficult or even impossible.

Containerising the application makes it possible to capture and preserve the software stack. It's not guaranteed to provide perfect reproducibility (for example, firmware and kernel versions may affect results) but it's a significant improvement on the current status quo.

- **Distributing validated HPC stacks to users:** The ease of sharing also makes it possible to HPC providers to provide containerised versions of their HPC software stacks. Users can then run on their own small clusters, workstations and laptops for development and initial debugging, for example.
- **Peer-to-peer sharing of stacks:** The ease of sharing stacks should also make it simpler to share applications between facilities in federated Grids (such as EGI) or even make it easier to balance workloads on different systems within a facility with less constraints to have a uniform software stack.
- **Enhancing cluster management and testing:** A completely different use case, but one worth mentioning, is to use containers on the management backend of HPC systems. For starters, one can quite easily simulate complex cluster stacks with a reasonable number of compute nodes on a laptop. Something that's not really possible with VMs. The logical next step would be to use containers in running production cluster services. There is no known site where this has been done yet though.

Containers vs. Virtualization

Many of the aforementioned cases can already be addressed in a virtualised HPC cloud environment, which are prevalent in many HPC centres today. However, there are some caveats:

- **Scheduling inefficiency:** With a container one can submit the jobs with your batch job scheduler (SLURM, PBS etc.) which has sophisticated queuing policy engines. With a VM you are typically dealing with a cloud management system (like OpenStack) which has a more basic scheduler that’s not designed for dispatching and queuing workloads in a HPC environment. Perhaps in the future advanced schedulers like Kubernetes and Mesos could accomplish this but this will likely require significant work.
- **Launch overhead:** The initialization of a VM takes a fairly long time compared to a container (<0.1 sec vs >20sec), so setting up an on-demand virtual cluster to run very short jobs carries a large overhead.
- **Resource overheads:** Running VMs can carry a larger overhead on both the amount of disk the images consume as well as the memory utilization.

There are certainly use cases where virtualisation remains relevant, for example running Windows instances, having long-running dedicated resources or wanting a fully isolated environment for security purposes to name a few.

Challenges

As described above, there are a number of benefits but also some challenges and questions remain.

Docker needs a fairly recent Linux distribution to work and privileged access to launch containers. These are addressed to a large extent by the aforementioned Shifter. However, it is still a product that’s very early in its development cycle and it has not been tested in real life yet (at least CSC from which we report experience here) so it remains to be seen how seamless it is. The security model of Docker is also being constantly improved and the development is proceeding at a rapid pace.

A second question that will be pertinent is how to deal with managing compatibility with low-level drivers (GPUs, parallel filesystems, interconnects etc.) which are exposed to the container. There will be some compatibility issues if, for example, a CUDA library is too new to the underlying kernel driver.

Another concern is the long term evolution of software in a containerized world. It is tempting to build a container manually but this will be difficult to manage in the long term. Ideally the build instructions should still be captured in some sort of configuration management framework and put under version control. Docker’s own Dockerfile format provides a fairly straightforward way for simple applications.

However, for more complex cases with many dependencies, using something more sophisticated and HPC-oriented like EasyBuild may be a good idea. An added benefit is that VMs and bare metal with the same configuration can also be targeted, as well as a variety of ready recipes for building various applications.

The licensing model of commercial applications, tools and libraries also poses a challenge. Which applications can be safely packaged into containers and shared in a repository and should the repository be protected for specific applications. The subject matter is complicated and poses both technical and legal questions.

It can be foreseen that there is a risk that containers are used too eagerly. Example: instead of putting an effort in how to adapt their application to the HPC-centre’s own typically well-defined and highly-tuned (bare metal) computing environment, the users will just deploy the application in the container in a “quick and dirty” fashion. Initially this may save work but in the long term this turns the application and its stack into a “black box”. This could result, for example, in performance issues and make debugging difficult.

The “container as magic bullet for everything” philosophy could also lead to reduction of investment into the “traditional” bare metal computing environment if all users are deploying their own containers with full stacks, resulting in stagnation of the standard environment. The worst case could be a cycle that ends up with all groups maintaining disparate stacks for their projects with performance typically lacking compared to the original, standard environment.

There are some potential practices to avoid the situation:

- Using containers judiciously, preferring “traditional” deployment into the bare metal computing environment if there’s no compelling reasons for containerising.
- Educating users to create “sustainable” containers that can be rebuilt and updated from Dockerfiles, EasyBuild blocks or some other configuration management system of their choice.

It would also help here to develop standard base containers for the HPC centres that are tuned to their systems, clean, well-maintained and compatible with the “bare metal” environment, having everything in configuration management. This is certainly an area where coordinated R&D effort and standardization work across PRACE partners and external entities such as XSEDE and would be beneficial.

Conclusion

Containers demonstrate significant potential in complementing bare metal computing environments and Virtual Machines and will almost certainly establish a strong standing in HPC in the next couple of years. However, there is still a considerable amount of work to do and potential caveats.

5 Management tools

This section provides information on HPC system management and monitoring software and systems currently in use by large HPC centres.

Critical issues on large-scale monitoring, from the perspectives of worldwide HPC centre system administrators, users, and vendors, would be to target, methodologies, desires, and data for gaining insights from large-scale system monitoring i.e. :

- identifying application performance variation causes;
- detecting contention for shared resources and assessing impacts;
- discovering misbehaving users and system components; and automating these analyses

Monitoring is vital so that resource usage can be billed to system owners, identify when systems (IT and plant) are getting into trouble, ensure security of systems from internet and physical attack (disaster) and ensure systems always fail-safe and alert the correct people.

Use cases of effective monitoring are the following:

- power/energy use
- failure analysis (potential for prediction)
- System level performance analysis (e.g. disabled components, interference contention)

Meaningful management tools are designed against well-defined requirements such as “What do we want to get from the data?” and “What has to be done to support this (collection, transport, analytics, storage, final use)?”

A key challenge is variability of quality of monitoring hardware and ‘query-ability’ from different vendors. Many systems have a very basic monitoring capability; in many cases handwritten instrumentation is needed.

Managing complex environments (hardware and software), could be a daunting task because of a heterogeneous hardware and system stack, multiple hybrid execution environments, converged HPC and extreme data workload management tools.

Big Challenges for effective monitoring can be categorised as follows:

- Scalability challenges:
    - hierarchical management architectures;
    - metrics and logs storing;
- Real time diagnostics:
    - apply data analytics methods to infrastructure management data;
    - high rate and accuracy measurement is not cheap;
    - provisioning sufficient bandwidth also can add cost;
- Complex workload resource allocation:
    - resource management (Multiple Programs / Multiple Data) MPMD applications;
    - flexibility (dynamic resources) and orchestration capabilities and
    - new workloads ecosystem (HPDA);
    - difficult to view larger data sets to extract insights.

5.1 System Monitoring

The following sections provide a list of relevant system monitoring tools used in various HPC centres worldwide

5.1.1 *CACTI*

Cacti [102] is a complete network graphing solution designed to harness the power of RRDTool's data storage and graphing functionality [103]. Cacti provides a fast poller, advanced graph templating, multiple data acquisition methods, and user management features out of the box. All of this is wrapped in an intuitive, easy to use interface that makes sense for LAN-sized installations up to complex networks with hundreds of devices.

Cacti is used to monitor Advanced Computing Facility ACF at EPCC. Cacti gathers all data from building management systems, metrics includes temperature, pressure, flow rate etc.

5.1.2 *OVIS*

OVIS [104] is a modular system for HPC data collection, transport, storage, analysis, visualization, and response. The OVIS project seeks to enable more effective use of High Performance Computational Clusters via greater understanding of applications' use of resources, including the effects of competition for shared resources; discovery of abnormal system conditions; and intelligent response to conditions of interest.

Data Collection, Transport, and Storage
The Lightweight Distributed Metric Service (LDMS) is the OVIS data collection and transport system. LDMS provides capabilities for lightweight run-time collection of high-fidelity data. Data can be accessed on-node or transported off node. Additionally, LDMS can store data in a variety of storage options.

Analysis and Visualization
OVIS includes 2 and 3D visual displays of deterministic information about state variables (e.g., temperature, CPU utilisation, fan speed), user-generated derived variables (e.g., aggregated memory errors over the lifespan of a job), and their aggregate statistics. Visual consideration of the cluster as a comparative ensemble, rather than singleton nodes, is a convenient and useful method for tuning cluster set-up and determining the effects of real-time changes in the cluster configuration and its environment.

5.1.3 *Total Recall*

User support personnel, systems engineers, and administrators of HPC installations need to be aware of log and telemetry information from different systems in order to perform routine tasks ranging from systems management to user inquiries. Total Recall [105] is an integrated, distributed HPC tailored monitoring system, based on a current generation software stack from the DevOps community, with integration into the work load management system. The goal of this system is to provide a quicker turnaround time for user inquiries in response to errors. Dashboards provide an overlay of system and node level events on top of correlated metrics data. This information is directly available for querying, manipulation, and filtering, allowing statistical analysis and aggregation of collected data. Furthermore, additional dashboards offer insight into how users are interacting with available resources and pin-point fluctuations in utilization. The system can integrate sources of information from other monitoring solutions and event-based sources.

In addition, generic system administration tasks, e.g. capacity planning, should be carried out by the systems engineering staff in cooperation with research groups, rather than by individual system users. Bottleneck identification, monitoring of environmental data such as energy consumption and hardware counters for voltages and temperatures of hundreds or more compute nodes, demands a capable and scalable system that can offer insight and perspective on all of these aspects of an HPC installation's lifetime and operating state. Data can be used to infer potentially failing components due to high stress and excessive strain encountered because of data-intensive applications executing in a demanding computing environment.

"Total Recall" goal is to achieve a uniform and user-friendly monitoring interface as a single entry point to these sources of information, and to offer consistent insights on all entity metrics in the monitored system. Specific details on system status and job execution can be prepared in separate dashboards for end-users (e.g. per-group and per-user) of the HPC system.

5.2 User Level Tools

5.2.1 *XDMOD (XD Metrics on Demand)*

XDMoD (XD Metrics on Demand [106]) is an NSF-funded open source tool designed to audit and facilitate the utilization of the XSEDE cyberinfrastructure by providing a wide range of metrics on XSEDE resources, including resource utilization, resource performance, and impact on scholarship and research. The XDMoD framework is designed to meet the following objectives: (1) provide the user community with a tool to manage their allocations and optimize their resource utilization, (2) provide operational staff with the ability to monitor and tune resource performance, (3) provide management with a tool to monitor utilization, user base, and performance of resources, and (4) provide metrics to help measure scientific impact. While initially focused on the XSEDE program, Open XDMoD has been created to be adaptable to any HPC environment.

The framework includes a computationally lightweight application kernel auditing system that utilizes performance kernels chosen from both low-level benchmarks and actual scientific and engineering applications to measure overall system performance from the user's perspective. This allows continuous resource monitoring to measure all aspects of system performance including file-system, processor, and memory performance, and network latency and bandwidth. Current and past utilization metrics, coupled with application kernel-based performance analysis, can be used to help guide future cyberinfrastructure investment decisions, plan system upgrades, tune machine performance, improve user job throughput, and facilitate routine system operation and maintenance.

5.2.2 *LLview: User-level System Monitoring*

LLview [107] is a client-server based application which allows to monitor the utilization of clusters controlled by batch systems like IBM LoadLeveler, PBSpro, Torque, or IBM Blue Gene system data base. On large supercomputer clusters with several thousands of processors it is not feasible to monitor the usage of system and batch load with command line tools, because the lists or tables in their output are becoming too large and complex. Important information for example is the usage of nodes/processors and the required resources of running and waiting jobs. LLview gives a quick and compact summary of this information.

It is not feasible to monitor the usage of system and batch load with command line tools, because the lists or tables in their output are becoming too large and complex. Therefore, the LLview monitoring application was developed and extended to monitor the utilization of these resources via graphical interface.

Graphical elements of LLview are a node display, a usage bar, which gives a direct view of the job-granularity, a list of running jobs and a list of waiting jobs, and a graph chart displaying the number of jobs in the different queues. A very flexible type of charts for displaying job profiles for I/O, memory and CPU usage over time is introduced with client version 1.41. Besides, generic annotations are integrated, which allow for arbitrarily placed texts and images to enrich LLview displays with additional information on any location of the screen.

In addition, administrators use LLview to get a load status overview of the computational resource they administrate. It represents a visualization of the mapping between running jobs and nodes of clusters controlled by a batch system. It offers a wide variety of illustrations in only one window, including efficient supervision node usage, running and waiting jobs, several statistics, a history of jobs as well as reservations. This fully configurable application provides interactive mouse sensitive information about resource usage as shown.

5.2.3 *XALT*

XALT [108] is a tool that allows supercomputer support staff to collect and understand job-level information about the libraries and executables that end-users access during their jobs. The tool can also work with a system's module software to provide additional information about module usage. XALT is a collaboration between PI Mark Fahey (University of Chicago, formerly National Institute for Computational Sciences) and co-PI Robert McLay (TACC).

Most computing centres need to answer the questions:

- How many users and projects use a particular library or executable?
- How many users use which compilers?
- Which centre provided packages are used often? And which ones are never used?
- Which users or applications still use old version of certain library, compiler, or executable?
- Are there any widely used user-installed packages that a centre should provide instead?

With the information that XALT collects, high-end computer administrators can answer our questions by tracking continuous job-level information to learn what products researchers do and do not need. Drilling down to data driven usage statistics helps stakeholders conduct business in a more efficient, effective, and systematic way.

5.2.4 *REMORA*

Knowing about the requirements of HPC applications is a common question that users of high performance systems ask often. However, answering this question requires the collaboration of administrators and sometimes the answer does not contain the amount of detail that users demand. This work introduces a new user space resource monitoring tool, REMORA [109]. REMORA stands for REsource MOnitoring for Remote Applications, and provides a simple interface to gather important system utilization data while running on HPC systems. REMORA is designed to provide a brief text report and post-processing tools to analyse the very detailed records taken during an application run. Users can configure the tool to achieve the amount of detail that they want and perform the analysis of the results at any point in time.

REMORA helps users to achieve a better understanding of their applications by providing a high level profile of their executions and users can take advantage of that information to improve their codes.

5.3 Energy Efficiency

5.3.1 *Redfish*

Designed to meet the expectations of end users for simple, modern and secure management of scalable platform hardware, the DMTF's Redfish is an open industry standard specification and schema that specifies a RESTful interface and utilizes JSON and OData to help customers integrate solutions within their existing tool chains.

The **Redfish Resource Explorer** [110] provides interactive mock-ups of Redfish implementations, showing typical data returned through the API, how that data is organized, and the definition of each property as an easy way to get familiar with the Redfish API.

5.3.2 *Power API*

Achieving practical exascale supercomputing will require massive increases in energy efficiency. The bulk of this improvement will likely be derived from hardware advances such as improved semiconductor device technologies and tighter integration, hopefully resulting in more energy efficient computer architectures. Still, software will have an important role to play. With every generation of new hardware, more power measurement and control capabilities are exposed. Many of these features require software involvement to maximize feature benefits. This trend will allow algorithm designers to add power and energy efficiency to their optimization criteria. Similarly, at the system level, opportunities now exist for energy-aware scheduling to meet external utility constraints such as time of day cost charging and power ramp rate limitations. Finally, future architectures might not be able to operate all components at full capability for a range of reasons including temperature considerations or power delivery limitations. Software will need to make appropriate choices about how to allocate the available power budget given many, sometimes conflicting considerations.

For these reasons, Sandia National Laboratories has developed a portable API for power measurement and control. This Power API [111] provides multiple levels of abstractions to satisfy the requirements of multiple types of users. The complete document specification can be found in [112].

5.4 Next-generation monitoring and log analytics – CSC's Experience

In recent years with the advent of web-scale services and cloud computing, the amount of large-scale clusters around the world has skyrocketed. Many of these new communities are rapidly developing sophisticated tooling for managing these systems. While many of these tools are not primarily targeted for high-performance computing, the challenges faced with managing and monitoring large clustered environments are fairly similar. A large part of CSC's monitoring renewal is based on exploring and leveraging these tools in a HPC setting.

Some criteria:

- Open source;

- Vendor-agnostic;
- Composable architecture;
- Scalable ;
- User friendly, modern interfaces;
- Can be leveraged also for IT monitoring beyond HPC.

While the toolchain consists of many tools, the most notable are the Monitoring and log analytics tools.

5.4.1 *Monitoring*

For monitoring, a combination of Collectd, Graphite and Grafana has replaced the venerable and classic RRDTool/Ganglia stack. It consists of the following main components:

- **Collectd** [113] for gathering node level statistics. The overhead is somewhat higher than the Ganglia collectors but the functionality is much richer.
- **Graphite** [114] for storing the database and creating simple graphs and CSV output.
- **Grafana** [115] for creating sophisticated dashboards from the Graphite monitoring data.

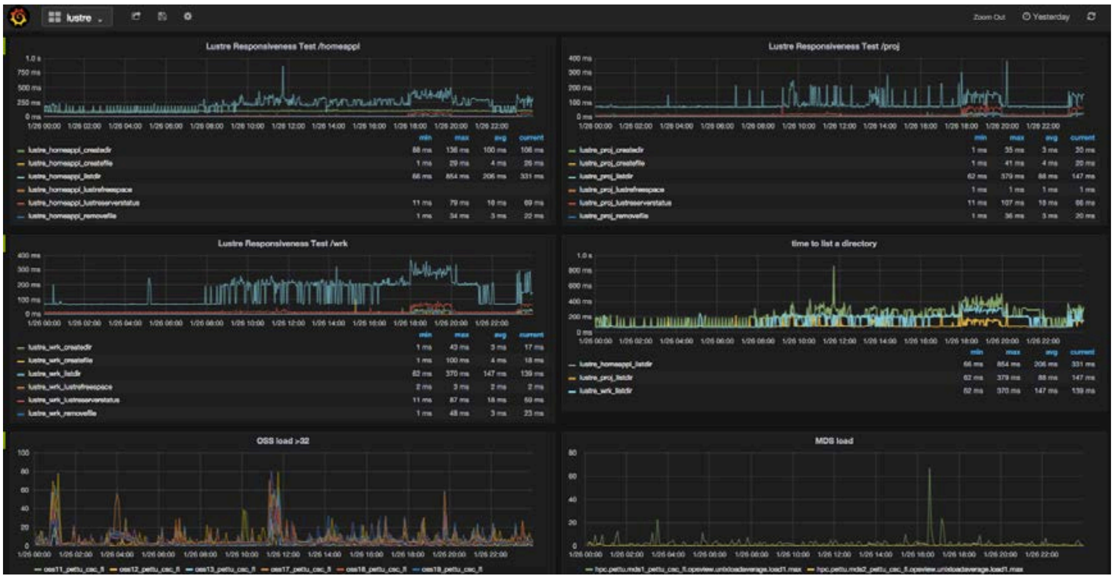

lustre
Zoom Out
Yesterday
Lustre Responsiveness Test /homeappl
1.0 s
750 ms
500 ms
250 ms
0 ms
1/26 00:00
1/26 02:00
1/26 04:00
1/26 06:00
1/26 08:00
1/26 10:00
1/26 12:00
1/26 14:00
1/26 16:00
1/26 18:00
1/26 20:00
1/26 22:00
min max avg current
lustre_homeappl_createdir 88 ms 136 ms 100 ms 106 ms
lustre_homeappl_createfile 1 ms 29 ms 4 ms 26 ms
lustre_homeappl_listdir 66 ms 854 ms 206 ms 331 ms
lustre_homeappl_lustrefreespace
lustre_homeappl_lustreserverstatus 11 ms 79 ms 18 ms 69 ms
lustre_homeappl_removefile 1 ms 34 ms 3 ms 22 ms
Lustre Responsiveness Test /proj
400 ms
300 ms
200 ms
100 ms
0 ms
min max avg current
lustre_proj_createdir 1 ms 35 ms 3 ms 20 ms
lustre_proj_createfile 1 ms 41 ms 4 ms 20 ms
lustre_proj_listdir 62 ms 379 ms 88 ms 147 ms
lustre_proj_lustrefreespace 1 ms 1 ms 1 ms 1 ms
lustre_proj_lustreserverstatus 11 ms 107 ms 18 ms 66 ms
lustre_proj_removefile 1 ms 36 ms 3 ms 20 ms
Lustre Responsiveness Test /wrk
400 ms
300 ms
200 ms
100 ms
0 ms
min max avg current
lustre_wrk_createdir 1 ms 43 ms 3 ms 17 ms
lustre_wrk_createfile 1 ms 100 ms 4 ms 18 ms
lustre_wrk_listdir 62 ms 370 ms 147 ms 139 ms
lustre_wrk_lustrefreespace 2 ms 3 ms 2 ms 2 ms
lustre_wrk_lustreserverstatus 11 ms 87 ms 18 ms 60 ms
lustre_wrk_removefile 1 ms 48 ms 3 ms 23 ms
time to list a directory
1.0 s
800 ms
600 ms
400 ms
200 ms
0 ms
min max avg current
lustre_homeappl_listdir 66 ms 854 ms 206 ms 331 ms
lustre_proj_listdir 62 ms 379 ms 88 ms 147 ms
lustre_wrk_listdir 62 ms 370 ms 147 ms 139 ms
OSS load >32
100
80
60
40
20
0
oss11_pettu_csc_fi
oss12_pettu_csc_fi
oss13_pettu_csc_fi
oss17_pettu_csc_fi
oss18_pettu_csc_fi
oss19_pettu_csc_fi
MDS load
80
60
40
20
0
hpc.pettu.mds1_pettu_csc_fi.opsview.unixloadaverage.load1.max
hpc.pettu.mds2_pettu_csc_fi.opsview.unixloadaverage.load1.max

Figure 55: **Example screenshot from Grafana of Lustre load monitoring**

There are a number of alternative tools that could be further used to extend and/or replace some components of the stack [116].

5.4.2 *Log analytics*

The amount of log files in a typical cluster can be significant and oftentimes an administrator needs to sift through a large amount of them, some example scenarios:

- Retroactively debugging a user-reported problem;
- Gaining rapid situational awareness in a technical or security incident (understanding the scope and root cause etc.)
- Gathering statistics (for example, compiler use);
- Proactively spotting anomalous behaviour.

Log analytics addresses these issues, and more. In this case, the “ELK stack” [117] was chosen. It is an open source combination of 3 different tools.

- **Logstash** for data collection and formatting;
- **Elasticsearch** for scalable, clustered data storage and searching;
- **Kibana** for visualization and data analysis.

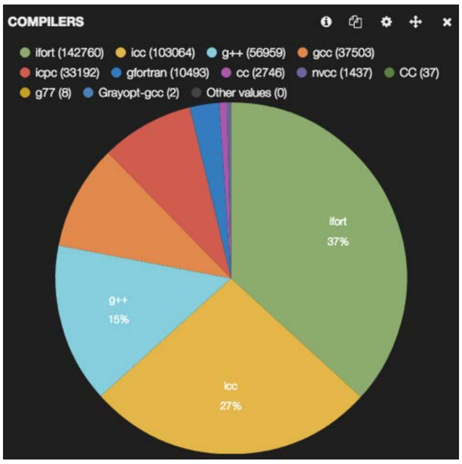


Figure 56: **Compiler invocation counts from CSC's Taito cluster frontends using ELK**

The stack has proven useful and is also rapidly being adopted by other groups at CSC and a end-customer pilot is in progress to provide Log analytics as a Service.

Combining the tools

It is also possible to interface the tools to each other. Some examples:

- Displaying ELK log events (such as job start/end) on the Grafana graphs using the overlay functionality;
- Generating Nagios alerts based on ELK events;
- Graphing ELK log information in Grafana.

5.4.3 *Experiences*

Building a stack with composable open source tools can be challenging and will require some dedicated work and skilled specialists. However, the resulting stack can provide very powerful tools.

Even as the tools have been set up, the development of dashboards and log analytics rules is an ongoing process with continuing improvement.

HPC systems should also be architected to facilitate the realtime shipment of logging and monitoring data out to these external collectors in a scalable way. With predesigned systems this may be sometimes difficult to achieve either due to the proprietary nature of the monitoring systems, performance limitations, or even both. System vendors should be actively encouraged to support highly scalable, standards-based APIs and message buses (such as RabbitMQ) to facilitate these dataflows.

5.4.4 *Other resources*

Similar toolchains are also being explored by the APEX consortium [118] as well as CERN and the HEPiX [119] community. A consulting company QNib Solutions, is also experimenting with combining a variety of monitoring and log sources with advanced toolchains [120].

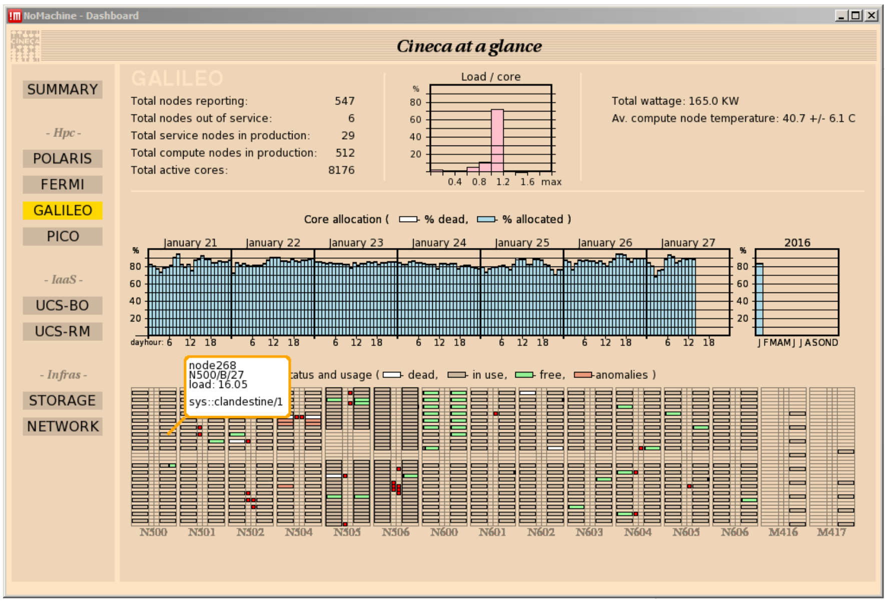
5.5 CINECA Dashboard – a single pane of glass for all data sources

The HPC systems of CINECA are constantly being monitored with the help of popular public domain and proprietary applications, such as Nagios, Ganglia, PBSPro, storage/network firmware and others. Each of these tools is providing a lot of useful information which helps to monitor the systems' health and promptly react in case of problems. However, the mere presence of multiple standard monitoring tools and large quantities of incoming data often makes it hard to take a global view of the system and decide on priority of interventions. Thus, the need to provide a bird's-eye view of the overall situation in the real time becomes quite pronounced.

To address this problem, CINECA formulated several principal requirements for a single-view dashboard application and started a corresponding internal project. The new tool is meant to offer a compact view of principal HPC systems and their ancillary infrastructure, show state of resource usage and allow for some of the standard real-time operations such as node draining/shutdown/restart and on-the-fly reorganization of batch queues.

The project started in the second half of 2015, and the first prototype application was implemented and tested by the end of the year for one of the clusters. Instead of developing a new applet/application for a browser which often is the choice for this type of tools, CINECA pointed at a combination of a customized remote NX desktop running a single program implemented in Tcl/Tk and hosted on a dedicated virtual machine. This combination has several advantages: NX allows for fast and secure access from any workplace, there is only one point of maintenance, and development cycle in for the Tcl/Tk interpreted language is short. As for the sources for incoming data, it was decided to involve only the existing monitoring agents and to not develop anything new at this stage. The data are being collected at the level of the cluster master nodes and then sent to the virtual host where Tcl/Tk application is running. This solution scales well, it allows for multiple simultaneously running desktops; each new desktop starts a copy of the application which occupies the entire NX window. Different users may be granted different views and privileges in accordance with their technical roles.

The project plans for 2016 include both adding of new functionalities and coverage of other HPC clusters.



Cineca at a glance

GALILEO

Total nodes reporting:	547
Total nodes out of service:	6
Total service nodes in production:	29
Total compute nodes in production:	512
Total active cores:	8176

Total wattage: 165.0 KW
Av. compute node temperature: 40.7 +/- 6.1 C

node268
N500/B/27
load: 16.05
sys::clandestine/1

Figure 57: A sample view of the current dashboard application

6 EU Projects for Exascale and Big Data

FP7 has funded exascale related efforts either via technology R&D projects (section 6.1) or via activities of PRACE IP projects (such as PRACE 3IP Pre-Commercial Procurement - PCP, see section 6.6).

Then H2020 amplified the programme, as explained in section 2.3.4), via FETHPC call of H2020 Work Programme 2014-2015 (now continued in Work Programme 2016-2017 [25][26]).

This chapter is giving a quick overview of these projects with brief technical contents hints, as well as on Centres of Excellence, plus some addenda on related or complementary efforts (Human Brain PCP, ITEA projects), and eventually on Coordination and Support Actions that support the European HPC ecosystem development. It shows a vivid research ecosystem that is gaining momentum under Horizon 2020.

Summaries of all FETHPC projects are collected in Annexe (section 8.1). More information can be found on line on H2020 portal, following the given references in this report [23] [131].

6.1 FP7

Deliverable D5.3 of PRACE 2IP [121] described the first exascale related EU projects: DEEP, Mont-Blanc and CRESTA as well as other projects started in 4Q2013: DEEP-ER, EPiGRAM, EXA2CT, NUMEXAS and Mont-Blanc 2 [122] (plus ITEA H4H reminded in section 6.5 below).

Below is only a very short reminder of those projects who are all finished or coming to an end. They are further mentioned in the synthetic view of section 6.2, together with FETHPC H2020 projects.

DEEP and DEEP-ER [123][124][125]
The DEEP concept uses a Cluster Booster Architecture as proof-of-concept for a next-generation 100 PFlop/s production system. Eurotech installed a prototype of the DEEP "Booster," a tightly coupled cluster of manycore coprocessors of 505 TFlop/s, at Jülich Supercomputing Centre, in July 2015. This novel and highly-scalable HPC system completes the Exascale-enabling DEEP System. The DEEP prototype system will remain in use at Jülich Supercomputing Centre at least for the next two years and also will be made available to application developers outside the project.

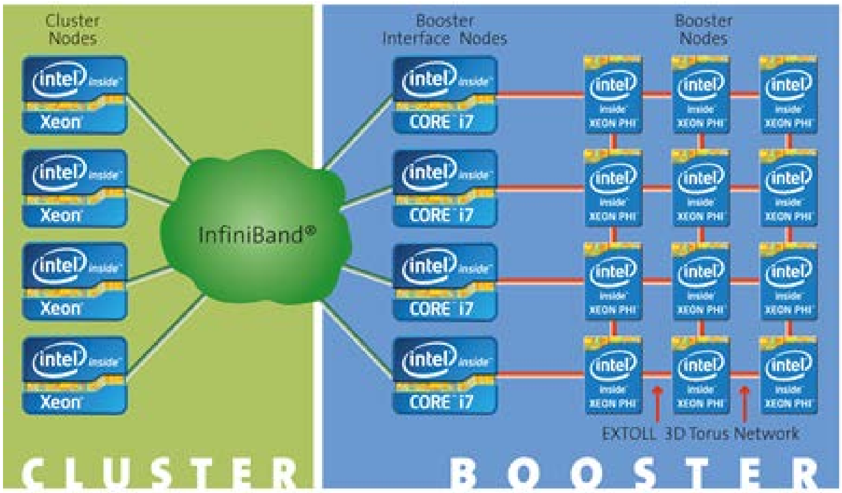


Figure 58: DEEP hardware architecture

DEEP extended Reach project (DEEP-ER [124]) extended the Cluster-Booster Architecture with a highly scalable, efficient, easy-to-use parallel I/O system and resiliency mechanisms.

Mont-Blanc/Mont-Blanc 2 [126]
The Mont-Blanc project started in October 2011 and until September 2014, followed by Mont-Blanc 2 (October 2011 – September 2016) – see also section 3.1.3. The Mont-Blanc project has set itself the 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 aimed at complementing Mont-Blanc with system software stack and evaluating ARM-based platforms.

CRESTA [127]
The CRESTA project (Collaborative Research into Exascale Systemware, Tools & Applications) was a FP7 EU project with the aim to provide exascale requirements from the end user point of view and new tools and systemware. The project had 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.

EPiGRAM [128]
The Exascale ProGRAmming Models (EPiGRAM,) project was an EC-funded FP7 project on exascale computing. The aim of the EPiGRAM project was to prepare Message Passing and PGAS programming models for exascale systems by fundamentally addressing their main current limitations.

EXA2CT [129]
The ***EXascale Algorithms and Advanced Computational Techniques*** project goal was 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.

NUMEXAS [130]
The NUMEXAS project (***Numerical Methods and Tools for Key Exascale Computing Challenges in Engineering and Applied Sciences***) started in October 2013 for 3 years.

A STREP collaborative project within the FP7-ICT programme of the European Union, the goal of Numexas was to develop, implement and validate the next generation of numerical methods running on exascale computing architectures.

6.2 H2020 – FETHPC

As mentioned in section 2.3.4, nineteen projects were selected in 2015 from H2020 Work Programme 2014-2015 FETHPC call, plus two Coordination and Support Actions (CSA) [131] – these latter ones being described below in section 6.4.

It is not possible to describe extensively all 19 technical projects – on-line documentation and factsheets can be found (H2020 portal) and the project summaries have been included in appendix (section 8.1).

Here we give a brief synthetic overview of the projects contents clustered by R&D topics.

The call was broken down into subtopics:

FETHPC 1 - 2014: HPC Core Technologies, Programming Environments and Algorithms for Extreme Parallelism and Extreme Data Applications
Scope: Proposals shall target one of the following subtopics:
A) HPC core technologies and architectures (e.g. processors, memory, interconnect and storage) and their optimal integration into HPC systems, platforms and prototypes
B) Programming methodologies, environments languages and tools: new programming models for extreme parallelism and extreme data applications
C) Application Programming Interfaces and system software for future extreme scale systems
D) New mathematical and algorithmic approaches for existing or emerging applications

Table 5: FETHPC 1-2015 H2020 call topic

The following table gives a first level classification of projects by main focus and topic w.r.t. this decomposition.

FETHPC1 Sub-topic addressed	Proposal Acronym
A - HPC core technologies and architectures 9 projects	ECOSCALE
	ExaNeSt
	ExaNoDe
	EXTRA
	greenFLASH
	MANGO
	Mont-Blanc 3
	NEXTGenIO
	SAGE
B - Programming methodologies, environments languages and tools 5 projects	ALLScale
	ANTAREX
	ESCAPE
	INTERTWINE
	READEX
D - New mathematical and algorithmic approaches 5 projects	ComPat
	ExaFLOW
	ExaHyPE
	ExCAPE
	NLAFET

Table 6: FETHPC 1-2015 H2020 call topic

In terms of funding, 70% of the ca. 90 M€ granted went to research organisations (academia and RTO), 15% to SMEs, 15% to (larger) industrial companies.

The lack of type C projects (Application Programming Interfaces and system software) was emphasized – ETP4HPC warning the European Commission that this was a shortcoming of the portfolio of projects, potentially jeopardising the balanced development of the Research Agenda.

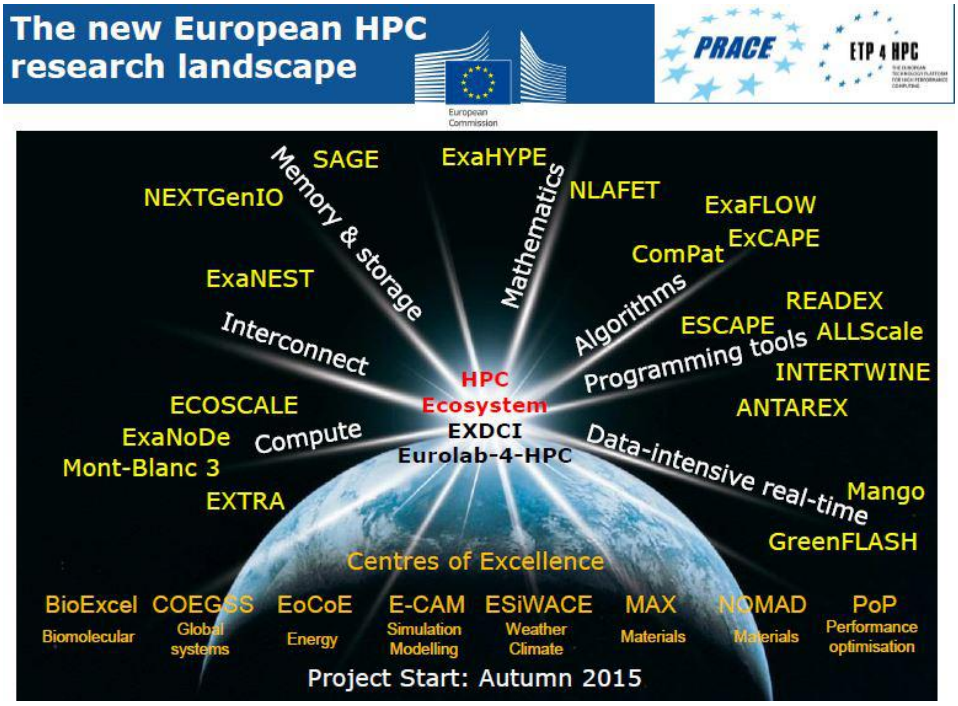


Figure 59: H2020 HPC projects (FETHPC and CoEs)

The different tables below gives a brief overview of technical areas and topics covered by the projects – as was presented during a SC15 BOF organised by ETP4HPC ([133]). This compilation, like “Figure 59: H2020 HPC projects (FETHPC and CoEs)” above, combines FETHPC projects and FP7 projects listed in section 6.1.

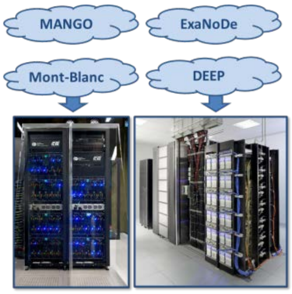
Energy Efficiency • Low power compute components • Advanced nanotechnologies • Extreme resource efficiency • Power monitoring • Power aware scheduling and programming • Hot-water cooling	Mont-Blanc DEEP MANGO ExaNoDe
Heterogeneity • Fine grained heterogeneity – Heterogeneous compute cores on chip – Mobile on chip GPUs • System-level modularity – Cluster-Booster approach – Custom interconnection • Hierarchical system partitioning – "Workers" grouped by address spaces	Mont-Blanc DEEP MANGO ECOSCALE
Reconfigurability • Support for specific HPC applications – Mapping of functions – Mapping of algorithms • Leveraging low reconfig. overhead • Aiming reducing data traffic	EXTRA ECOSCALE
Balanced / Co-design driven • Heterogeneous hardware platform with matching software stack and optimized grand-challenge HPC applications • Analysis of requirements HPC applications, mini-apps and kernels via performance analysis tools • Programming models enabling hw resources with minimal impact on applications	DEEP Mont-Blanc EXTRA ExaNoDe
Integration and Reliability • High density – 2.5D technology / Interposer – System on Package • Advanced cooling – Air/Liquid – Single/double loop – Heat reuse • Fault tolerance – Error detection mechanisms – Reliability of large systems – Quality of Service	

Table 7: Architecture and Compute topics of EU HPC projects

Interconnect and Memory • Develop a new server architecture using next generation interconnection, and memory advances – Integration of NVRAM technologies in the I/O stack – Fast, distributed in-node non-volatile memory Storage – Extreme compute power density – Low-latency unified Interconnect for compute & storage traffic – Codesign using accelerators and FPGAs – Liquid cooling technologies	NextGENIO SAGE ExaNEST
Storage • Develop the systemware to support new architectures use at the Exascale – Data Centric Computing System based on object-storage – Leveraging the I/O stack to enhance data-intensive computing – Computation off-loading to I/O system • Model different I/O workloads and use this understanding in a co-design process – Very Tightly Coupled Data & Computation – API for massive data ingest and extreme I/O – Extreme data management and analysis	SAGE NextGENIO
Data-Intensive RTS • Exploring New Heterogeneous Architectures For HPC Systems – Deeply heterogeneous manycore architectures – Real-time support providing a unified access to the systems via a smart interconnect – Adaptive programming models and compiler support to the new architectures – New applications and use cases emerging for HPC arena – Real-time, data-intensive, and energy efficiency – Applications used to identify system requirements – Real scientific and data center applications (e.g. MAORY RTC system).	MANGO Green FLASH

Table 8: Interconnect, Memory & Storage Data-Intensive Real Time topics of EU HPC projects

PROGRAMMING MODELS	• Innovative Programming Models for Exascale – **AllScale**: An exascale programming, multi-objective optimization and resilience management system supporting nested recursive parallelism. • Enhanced MPI and PGAS: the incremental approach – **CRESTA**: Enhancing programming models and system software by co-design. Large-scale real-world applications to guide the development of the software stack for exascale. – **EPiGRAM**: MPI and GPI for exascale. Combing best features of MessagePassing and PGAS programming model. – **Exa2CT**: cutting edge of the development of solvers, related algorithmic techniques, and HPC software architects for GASPI communication.
INTEROPERABILITY AND AUTOTUNING	• Interoperability of programming models: the ”+” issue – **Intertwine**: enhanced programming and runtime systems for effective interoperability. • Autotuning for energy efficiency and green HPC: – **READEX**: developing a tools-aided methodology for dynamic auto-tuning of HPC applications to exploit the dynamically changing resource requirements for improved energy-efficiency. – **ANTAREX**: providing a breakthrough approach to express by a DSL the application of self-adaptivity and to runtime manage and autotune applications for green and heterogeneous HPC.
ALGORITHMS	• Computational Fluid Dynamics for Exascale – **ExaFLOW**: addressing algorithmic challenges to enable the use of accurate simulation models on exascale with focus on error control, AMR in complex geometries, heterogenous modeling, energy efficiency, and in-site I/O. – **NUMEXAS**: develop, implement and demonstrate the next generation of numerical simulation methods with focus on industrial applications and on pre- and post-processing – **ESCAPE**: developing next generation IFS numerical blocks for weather forecast. • Multiscale Applications – **ComPAT**: development of the High Performance Multiscale Computing paradigm.
ALGORITHMS AND MATHEMATICS	• Machine Learning: – **ExCAPE**: better machine learning algorithms for predicting biological activity of drugs and their deployment on HPC systems. • Solvers: – **NLAFET**: linear solvers for exascale with novel algorithms, advanced scheduling strategies and autotuning – **ExaHyPE**: High-order Discontinous Galerkin hyperbolic PDE engine for geo- and astrophysics.

Table 9: Programming Models, Algorithms and Mathematics topics of EU HP projects

6.3 Centres of Excellence for Computing Applications

Eight new Centres of Excellence (CoEs) for computing applications have been selected following the recent call under e-Infrastructures [134]. They will help strengthen Europe's existing leadership in HPC (high-performance computing) applications and cover important areas like renewable energy, materials modelling and design, molecular and atomic modelling, climate change, Global System science, and bio-molecular research, and tools to improve HPC applications performance.

The projects retained are:

- EoCoE - Energy oriented Centre of Excellence for computer applications;
- BioExcel - Centre of Excellence for Biomolecular Research;
- NoMaD - The Novel Materials Discovery Laboratory;
- MaX - Materials design at the eXascale;
- ESiWACE - Excellence in SImulation of Weather and Climate in Europe;
- E-CAM - An e-infrastructure for software, training and consultancy in simulation and modelling;
- POP - Performance Optimisation and Productivity;
- COEGSS - Center of Excellence for Global Systems Science.

6.4 Coordination and support actions: EXDCI and Eurolab4HPC - NESUS

EXDCI and Eurolab4HPC were funded under FETHPC-1 call in Work Programme 2014-205 of H2020.

6.4.1 *EXDCI*

The European Extreme Data & Computing Initiative (EXDCI) objective is to coordinate the development and implementation of a common strategy for the European HPC Ecosystem [135]. The two most significant HPC bodies in Europe, PRACE and ETP4HPC, join their expertise in this 30-month project with a budget of € 2.5 million, starting from September 2015. EXDCI aims to support the road-mapping, strategy-making and performance-monitoring activities of the ecosystem, i.e.:

- Producing and aligning roadmaps for HPC Technology and HPC Applications;
- Measuring the implementation of the European HPC strategy;
- Building and maintaining relations with other international HPC activities and regions;
- Supporting the generation of young talent as a crucial element of the development of European HPC.

EXDCI will complement the Horizon 2020 calls and projects in the achievement of a globally competitive HPC Ecosystem in Europe. Following the vision of the European Commission in HPC, this ecosystem is based on three pillars: HPC Technology Provision, HPC Infrastructure and HPC Application Resources.

6.4.2 *Eurolab4HPC*

EuroLab-4-HPC is a two-year Horizon 2020 funded project with the bold commitment to build the foundation for a European Research Center of Excellence in High-Performance Computing (HPC) Systems [136]. The main objectives are:

- to join HPC system research groups around a long-term HPC research agenda by forming an HPC research roadmap and joining forces behind it;
- to define an HPC curriculum in HPC technologies and best-practice education/training methods to foster future European technology leaders;
- to accelerate commercial uptake of new HPC technologies;
- to build an HPC ecosystem with researchers and other stakeholders, e.g., HPC system providers and venture capital;
- to form a business model and organization for the EuroLab-4-HPC excellence centre in HPC systems.

EuroLab-4-HPC is coordinated by Chalmers University of Technology and involves thirteen prominent research organizations across nine countries with some of the best research teams in HPC in Europe.

6.4.3 *NESUS*

COST – European Cooperation in Science and Technology is an intergovernmental framework aimed at facilitating the collaboration and networking of scientists and researchers at the European level. NESUS is an EU-funded COST action devoted to sustainability in ultrascale systems, started in April 2014 [137]. It has work groups covering aspects to make systems more usable and sustainable: programming models, resilience, runtime systems, data management, and energy efficiency. NESUS also focusses on cross cutting topics such as training and SMEs, and fosters international cooperation between scientists and industry.

6.5 Others HPC technology R&D projects: ITEA programmes

Not part of H2020, ITEA is the EUREKA Cluster programme supporting innovative, industry-driven, pre-competitive R&D projects in the area of Software-intensive Systems & Services (SiSS). Funding mechanisms and rules are different than H2020, relying on specific member states involvement.

ITEA stimulates projects in an open community of large industry, SMEs, universities, research institutes and user organisations. Each year, ITEA issues a Call for projects starting with a two-day brokerage event. Each Call follows a two-stage procedure, in which the quality of the project proposal is evaluated and improved, finally leading to a selection of high quality project proposals that receive the official ITEA label.

ITEA successive programmes (since 2008) supported various HPC projects with a good industry/research&academia intertwining [138], briefly mentioned in Table 10 below.

ITEA 2 Call 1 ParMA	Parallel Programming for Multi-core Architectures	Technologies to exploit the power of multicore architectures
ITEA 2 Call 2 HiPiP	High Performance image Processing	Reducing complex image-processing latency to improve immediate use of image information.
ITEA 2 Call 4 H4H	Hybrid4HPC	Advanced tools and application framework for future technologies and HPC architectures
ITEA 2 Call 8 COLOC	COncurrency and LOcality Challenge	Methodologies and tools for data placement (resource allocation and access) and thread concurrency management.

Table 10: ITEA HPC projects

6.6 Pre-Commercial Procurements (PRACE, HBP)

PCP [139] is an instrument promoted by the European Commission (EC) to foster innovation through public procurement. It allows to procure research and development services to enable development of new solutions which would otherwise likely not be available. By design a PCP is organized as a multi-phase, competitive process. During phase I suppliers, which have been awarded a contract, should work on solution designs. Typically the number of contracts awarded in the consecutive phases is becoming smaller. During phase II contractors work on prototypes to prepare for small product/service development in the final phase III, which includes field tests in relevant environment for a system at TRL 8 or 9 [141].

6.6.1 *PRACE*

The goal of the PCP carried out by a Group of Procurers within PRACE-3IP is to facilitate whole system design for energy efficient HPC that should lead to HPC solutions, which on the one hand are suitable for operation within the PRACE infrastructure of leadership class systems for scientific computing, and on the other hand significantly improves on energy efficiency. The bidders were given the freedom to propose different solutions with respect to how to achieve improvements in terms of energy efficiency. These improvements in energy efficiency must be demonstrated through the use of real production application codes and a subset of benchmark suite in use by PRACE (the Unified European Applications Benchmark Suite UEABS [142]).

The PCP opened a call for tender in November 2013. Currently the PCP is in phase II, for which the following suppliers had been awarded with a contract: Bull SAS (France), E4 Computer Engineering (Italy) and Maxeler Technologies (UK). During the following phase III, which is expected to start in summer 2016, the remaining 2 contractors will have to deploy pilot system with a compute capability of around 1 PFlop/s, to demonstrate technology readiness of the proposed solution and the progress in terms of energy efficiency. NB: this PCP topic is echoing FETHCP Energy Efficiency topic mentioned in Table 7.

6.6.2 *Human Brain Project*

By developing and expanding the use of information technology, the Human Brain Project (HBP, [140]) aims for opening new opportunities for brain research. Particular challenging is the enablement of large-scale simulations of brain models as today’s HPC architectures do not meet their requirements. This includes both, the need for extremely large memory footprint and interactive supercomputing. For realistic network sizes the amount of data generated during a simulation becomes too large for being written to an external storage system and the complexity of the simulations requires interactive steering.

To ensure that suitable solutions for realizing HBP's future High-Performance Analytics and Computing Platform will exist, the project published in April 2014 a tender for a PCP focussing on R&D services in the following areas: integration of dense memory technologies, scalable visualization as well as dynamic management of resources required for interactive access to the systems.

At the time of writing this report, the PCP was in phase III, for which Cray and a consortium consisting of IBM and NVIDIA had been awarded a contract. These contractors are implementing their proposed solutions and demonstrate their technological readiness on pilot systems that will be installed later in 2016.

7 Conclusion

This first Technology Watch deliverable of PRACE-4IP Work Package 5 gives an updated overview of HPC and Big Data trends, in terms of technologies and with some market and business analysis hints. It uses a diversity of sources: Top500 and publicly available market data and analyses, recent supercomputing conferences (mostly SC15 for this current report), other HPC events and public literature, direct (non NDA) contacts with vendors and direct participation of WP5 members in a diversity of European projects or initiatives. Technical, as well as operational aspects, related to computing infrastructures are further investigated in Task 2, with also corresponding best practices for the design and commissioning of HPC facilities. Best practices regarding prototyping and technology assessment are dealt with in Task 3. The combination of these three tasks outcomes make up a consistent and living portfolio of practical documentation for insight and guidance in the area of “Best Practices for HPC Systems Commissioning”.

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

- Chapter 2: “Worldwide HPC landscape and market overview” gave an updated as well as dynamic view of global trends – technology paths, vendors momentum, strategies and large initiatives by continent;
- Chapter 3: “Core technologies and components” gave a quick overview of processors, accelerators, memory and storage technologies, interconnect technologies;
- Chapter 4: “Solution and architectures” gave quick vendor snapshots, with some more specific trends and illustrations regarding storage, cooling and virtualisation and cloud delivery;
- Chapter 5: “Management tools” was an overview of various tools at system, user resources or log analysis levels;
- Chapter 6: “EU Projects for Exascale and Big Data” gave a more detailed view of EU activities that were briefly introduced in Chapter 2, and shows the excellent European momentum in the area of HPC technologies.

This deliverable will be followed by another similar report in one year.

8 Annex

8.1 Summaries of 19 FETHPC projects (Work Programme 2014-2015 of Horizon 2020)

ExaNoDe - European Exascale Processor Memory Node Design

ExaNoDe will develop and pilot (technology readiness level 7) a highly efficient, highly integrated, multi-way, high-performance, heterogeneous compute element aimed towards exascale computing and demonstrated using hardware-emulated interconnect. It will build on multiple European initiatives for scalable computing, utilizing low-power processors and advanced nanotechnologies. ExaNoDe will draw heavily on the Unimem* memory and system design paradigm defined within the EUROSERVER FP7 project, providing low-latency, high-bandwidth and resilient memory access, scalable to Exabyte levels. The ExaNoDe compute element aims towards exascale compute goals through:

- Integration of the most advanced low-power processors and accelerators across scalar, SIMD, GPGPU and FPGA processing elements supported by research and innovation in the deployment of associated nanotechnologies and in the mechanical requirements to enable the development of a high-density, high-performance integrated compute element with advanced thermal characteristics and connectivity to the next generation of system interconnect and storage;
- Undertaking essential research to ensure the ExaNoDe compute element provides necessary support of HPC applications including I/O and storage virtualization techniques, operating system and semantically aware runtime capabilities and PGAS, OpenMP and MPI paradigms;
- The development an instantiation of a hardware emulation of interconnect to enable the evaluation of Unimem for the deployment of multiple compute elements and the evaluation, tuning and analysis of HPC mini-apps. Each aspect of ExaNoDE is aligned with the goals of the ETP4HPC. The work will be steered by first-hand experience and analysis of high-performance applications, their requirements and the tuning of their kernels.

ExaNeSt - European Exascale System Interconnect and Storage

ExaNeSt will develop, evaluate, and prototype the physical platform and architectural solution for a unified Communication and Storage Interconnect and the physical rack and environmental structures required to deliver European Exascale Systems. The consortium brings technology, skills, and knowledge across the entire value chain from computing IP to packaging and system deployment; and from operating systems, storage, and communication to HPC with big data management, algorithms, applications, and frameworks. Building on a decade of advanced R&D, ExaNeSt will deliver the solution that can support exascale deployment in the follow-up industrial commercialization phases. Using direction from the ETP4HPC roadmap and soon-available high density and efficiency compute, we will model, simulate, and validate through prototype, a system with:

1. High throughput, low latency connectivity, suitable for exascale-level compute, their storage, and I/O, with congestion mitigation, QoS guarantees, and resilience.
2. Support for distributed storage located with the compute elements providing low latency that nonvolatile memories require, while reducing energy, complexity, and costs.
3. Support for task-to-data sw locality models to ensure minimum data communication energy overheads and property maintenance in databases.
4. Hyper-density system integration scheme that will develop a modular, commercial, European-sourced advanced cooling system for exascale in ~200 racks while maintaining reliability and cost of ownership.
5. The platform management scheme for big-data I/O to this resilient, unified distributed storage compute architecture.
6. Demonstrate the applicability of the platform for the complete spectrum of Big Data applications, e.g. from HPC simulations to Business Intelligence support. All aspects will be steered and validated with the first-hand experience of HPC applications and experts, through kernel turning and subsequent data management and application analysis.

NEXTGenIO - Next Generation I/O for Exascale

The overall objective of the Next Generation I/O Project (NEXTGenIO) is to design and prototype a new, scalable, high-performance, energy efficient computing platform designed to address the challenge of delivering scalable I/O performance to applications at the Exascale. It will achieve this using highly innovative, non-volatile, dual in-line memory modules (NV-DIMMs). These hardware and systemware developments will be coupled to a co-design approach driven by the needs of some of today’s most demanding HPC applications. By meeting this overall objective, NEXTGenIO will solve a key part of the Exascale challenge and enable HPC and Big Data applications to overcome the limitations of today’s HPC I/O subsystems. Today most high-end HPC systems employ data storage separate from the main system and the I/O subsystem often struggles to deal with the degree of parallelism present. As we move into the domain of extreme parallelism at the Exascale we need to address I/O if such systems are to deliver appropriate performance and efficiency for their application user communities. The NEXTGenIO project will explore the use of NV-DIMMs and associated systemware developments through a co-design process with three ‘end-user’ partners: a high-end academic HPC service provider, a numerical weather forecasting service provider and a commercial on-demand HPC service provider. These partners will develop a set of I/O workload simulators to allow quantitative improvements in I/O performance to be directly measured on the new system in a variety of research configurations. Systemware software developed in the project will include performance analysis tools, improved job schedulers that take into account data locality and energy efficiency, optimised programming models, and APIs and drivers for optimal use of the new I/O hierarchy. The project will deliver immediately exploitable hardware and software results and show how to deliver high performance I/O at the Exascale.

Mont-Blanc 3 - Mont-Blanc 3, European scalable and power efficient HPC platform based on low-power embedded technology

The main target of the Mont-Blanc 3 project "European Scalable and power efficient HPC platform based on low-power embedded technology" is the creation of a new high-end HPC platform (SoC and node) that is able to deliver a new level of performance / energy ratio whilst executing real applications. The technical objectives are:

1. To design a well-balanced architecture and to deliver the design for an ARM based SoC or SoP (System on Package) capable of providing pre-exascale performance when implemented in the time frame of 2019-2020. The predicted performance target must be measured using real HPC applications.
2. To maximise the benefit for HPC applications with new high-performance ARM processors and throughput-oriented compute accelerators designed to work together within the well-balanced architecture .
3. To develop the necessary software ecosystem for the future SoC. This additional objective is important to maximize the impact of the project and make sure that this ARM architecture path will be successful in the market. The project shall build upon the previous Mont-Blanc & Mont-Blanc 2 FP7 projects, with ARM, BSC & Bull being involved in Mont-Blanc 1, 2 and 3 projects. It will adopt a co-design approach to make sure that the hardware and system innovations are readily translated into benefits for HPC applications. This approach shall integrate architecture work (WP3 & 4 - on balanced architecture and computing efficiency) together with a simulation work (to feed and validate the architecture studies) and work on the needed software ecosystem.

SAGE

Worldwide data volumes are exploding and islands of storage remote from compute will not scale. We will demonstrate the first instance of intelligent data storage, uniting data processing and storage as two sides of the same rich computational model. This will enable sophisticated, intention-aware data processing to be integrated within a storage systems infrastructure, combined with the potential for Exabyte scale deployment in future generations of extreme scale HPC systems. Enabling only the salient data to flow in and out of compute nodes, from a sea of devices spanning next generation solid state to low performance disc we enable a vision of a new model of highly efficient and effective HPC and Big Data demonstrated through the SAGE project. Objectives

- Provide a next-generation multi-tiered object-based data storage system (hardware and enabling software) supporting future-generation multi-tier persistent storage media supporting integral computational capability, within a hierarchy.
- Significantly improve overall scientific output through advancements in systemic data access performance and drastically reduced data movements.
- Provides a roadmap of technologies supporting data access for both Exascale/Exabyte and High Performance Data Analytics.
- Provide programming models, access methods and support tools validating their usability, including ‘Big-Data’ access and analysis methods
- Co-Designing and validating on a smaller representative system with earth sciences, meteorology, clean energy, and physics communities
- Projecting suitability for extreme scaling through simulation based on evaluation results.

Call Alignment: We address storage data access with optimized systems for converged Big Data and HPC use, in a co-design process with scientific partners and applications from many domains. System effectiveness and power efficiency are dramatically improved through minimized data transfer, with extreme scaling and resilience.

MANGO: exploring Manycore Architectures for Next-GeneratiOn HPC systems

MANGO targets to achieve extreme resource efficiency in future QoS-sensitive HPC through ambitious cross-boundary architecture exploration for performance/power/predictability (PPP) based on the definition of new-generation high-performance, power-efficient, heterogeneous architectures with native mechanisms for isolation and quality-of-service, and an innovative two-phase passive cooling system. Its disruptive approach will involve many interrelated mechanisms at various architectural levels, including heterogeneous computing cores, memory architectures, interconnects, run-time resource management, power monitoring and cooling, to the programming models. The system architecture will be inherently heterogeneous as an enabler for efficiency and application-based customization, where general-purpose compute nodes (GN) are intertwined with heterogeneous acceleration nodes (HN), linked by an across-boundary homogeneous interconnect. It will provide guarantees for predictability, bandwidth and latency for the whole HN node infrastructure, allowing dynamic adaptation to applications. MANGO will develop a toolset for PPP and explore holistic pro-active thermal and power management for energy optimization including chip, board and rack cooling levels, creating a hitherto inexistent link between HW and SW effects at all layers. Project will build an effective large-scale emulation platform. The architecture will be validated through noticeable examples of application with QoS and high-performance requirements. Ultimately, the combined interplay of the multi-level innovative solutions brought by MANGO will result in a new positioning in the PPP space, ensuring sustainable performance as high as 100 PFLOPS for the realistic levels of power consumption (<15MWatt) delivered to QoS-sensitive applications in largescale capacity computing scenarios providing essential building blocks at the architectural level enabling the full realization of the ETP4HPC strategic research agenda

ECOSCALE - Energy-efficient Heterogeneous COmputing at exaSCALE

In order to reach exascale performance current HPC servers need to be improved. Simple scaling is not a feasible solution due to the increasing utility costs and power consumption limitations. Apart from improvements in implementation technology, what is needed is to refine the HPC application development as well as the architecture of the future HPC systems. ECOSCALE tackles this challenge by proposing a scalable programming environment and hardware architecture tailored to the characteristics and trends of current and future HPC applications, reducing significantly the data traffic as well as the energy consumption and delays. We first propose a novel heterogeneous energy-efficient hierarchical architecture and a hybrid MPI+OpenCL programming environment and runtime system. The proposed architecture, programming model and runtime system follows a hierarchical approach where the system is partitioned into multiple autonomous Workers (i.e. compute nodes). Workers are interconnected in a tree-like structure in order to form larger Partitioned Global Address Space (PGAS) partitions, which are further hierarchically interconnected via an MPI protocol. Secondly, to further increase the energy efficiency of the system as well as its resilience, the Workers will employ reconfigurable accelerators that can perform coherent memory accesses in the virtual address space utilizing an IOMMU. The ECOSCALE architecture will support shared partitioned reconfigurable resources accessed by any Worker in a PGAS partition, and, more importantly, automated hardware synthesis of these resources from an OpenCL-based programming model. We follow a co-design approach that spans a scalable HPC hardware platform, a middleware layer, a programming and a runtime environment as well as a high-level design environment for mapping applications onto the system. A proof of concept prototype and a simulator will be built in order to run two real-world HPC applications and several benchmarks.

EXTRA - Exploiting eXascale Technology with Reconfigurable Architectures

To handle the stringent performance requirements of future exascale High Performance Computing (HPC) applications, HPC systems need ultraefficient heterogeneous compute nodes. To reduce power and increase performance, such compute nodes will require reconfiguration as an intrinsic feature, so that specific HPC application features can be optimally accelerated at all times, even if they regularly change over time. In the EXTRA project, we create a new and flexible exploration platform for developing reconfigurable architectures, design tools and HPC applications with run-time reconfiguration built-in from the start. The idea is to enable the efficient co-design and joint optimization of architecture, tools, applications, and reconfiguration technology in order to prepare for the necessary HPC hardware nodes of the future. The project EXTRA covers the complete chain from architecture up to the application:

- More coarse-grain reconfigurable architectures that allow reconfiguration on higher functionality levels and therefore provide much faster reconfiguration than at the bit level.
- The development of just-in time synthesis tools that are optimized for fast (but still efficient) re-synthesis of application phases to new, specialized implementations through reconfiguration.
- The optimization of applications that maximally exploit reconfiguration.
- Suggestions for improvements to reconfigurable technologies to enable the proposed reconfiguration of the architectures.

In conclusion, EXTRA focuses on the fundamental building blocks for run-time reconfigurable exascale HPC systems: new reconfigurable architectures with very low reconfiguration overhead, new tools that truly take reconfiguration as a design concept, and applications that are tuned to maximally exploit run-time reconfiguration techniques. Our goal is to provide the European platform for run-time reconfiguration to maintain Europe’s competitive edge and leadership in run-time reconfigurable computing.

ESCAPE - Energy-efficient SCalable Algorithms for weather Prediction at Exascale

ESCAPE will develop world-class, extreme-scale computing capabilities for European operational numerical weather prediction (NWP) and future climate models. The biggest challenge for state-of-the-art NWP arises from the need to simulate complex physical phenomena within tight production schedules. Existing extreme-scale application software of weather and climate services is ill-equipped to adapt to the rapidly evolving hardware. This is exacerbated by other drivers for hardware development, with processor arrangements not necessarily optimal for weather and climate simulations. ESCAPE will redress this imbalance through innovation actions that fundamentally reform Earth-system modelling. ESCAPE addresses the ETP4HPC SRA 'Energy and resiliency' priority topic, developing a holistic understanding of energy-efficiency for extreme-scale applications using heterogeneous architectures, accelerators and special compute units. The three key reasons why this proposal will provide the necessary means to take a huge step forward in weather and climate modelling as well as interdisciplinary research on energy-efficient highperformance computing are: 1) Defining and encapsulating the fundamental algorithmic building blocks ("Weather & Climate

Dwarfs") underlying weather and climate services. This is the pre-requisite for any subsequent co-design, optimization, and adaptation efforts. 2) Combining groundbreaking frontier research on algorithm development for use in extreme-scale, high-performance computing applications, minimizing time- and cost-to-solution. 3) Synthesizing the complementary skills of all project partners. This includes ECMWF, the world leader in global NWP together with leading European regional forecasting consortia, teaming up with excellent university research and experienced high-performance computing centres, two world-leading hardware companies, and one European start-up SME, providing entirely new knowledge and technology to the field.

ComPat - Computing Patterns for High Performance Multiscale Computing

Multiscale phenomena are ubiquitous and they are the key to understanding the complexity of our world. Despite the significant progress achieved through computer simulations over the last decades, we are still limited in our capability to accurately and reliably simulate hierarchies of interacting multiscale physical processes that span a wide range of time and length scales, thus quickly reaching the limits of contemporary high performance computing at the tera- and petascale. Exascale supercomputers promise to lift this limitation, and in this project we will develop multiscale computing algorithms capable of producing high-fidelity scientific results and scalable to exascale computing systems. Our main objective is to develop generic and reusable High Performance Multiscale Computing algorithms that will address the exascale challenges posed by heterogeneous architectures and will enable us to run multiscale applications with extreme data requirements while achieving scalability, robustness, resiliency, and energy efficiency. Our approach is based on generic multiscale computing patterns that allow us to implement customized algorithms to optimise load balancing, data handling, fault tolerance and energy consumption under generic exascale application scenarios. We will realise an experimental execution environment on our pan-European facility, which will be used to measure performance characteristics and develop models that can provide reliable performance predictions for emerging and future exascale architectures. The viability of our approach will be demonstrated by implementing nine grand challenge applications which are exascale-ready and pave the road to unprecedented scientific discoveries. Our ambition is to establish new standards for multiscale computing at exascale, and provision a robust and reliable software technology stack that empowers multiscale modellers to transform computer simulations into predictive science.

ExCAPE - Exascale Compound Activity Prediction Engine

Scalable machine learning of complex models on extreme data will be an important industrial application of exascale computers. In this project, we take the example of predicting compound bioactivity for the pharmaceutical industry, an important sector for Europe for employment, income, and solving the problems of an ageing society. Small scale approaches to machine learning have already been trialed and show great promise to reduce empirical testing costs by acting as a virtual screen to filter out tests unlikely to work. However, it is not yet possible to use all available data to make the best possible models, as algorithms (and their implementations) capable of learning the best models do not scale to such sizes and heterogeneity of input data. There are also further challenges including imbalanced data, confidence estimation, data standards model quality and feature diversity. The ExCAPE project aims to solve these problems by producing state of the art scalable algorithms and implementations thereof suitable for running on future Exascale machines.

These approaches will scale programs for complex pharmaceutical workloads to input data sets at industry scale. The programs will be targeted at exascale platforms by using a mix of HPC programming techniques, advanced platform simulation for tuning and and suitable accelerators.

NLAFET - Parallel Numerical Linear Algebra for Future Extreme-Scale Systems

The NLAFET proposal is a direct response to the demands for new mathematical and algorithmic approaches for applications on extreme scale systems, as identified in the FETHPC work programme and call. This project will enable a radical improvement in the performance and scalability of a wide range of real-world applications relying on linear algebra software, by developing novel architecture-aware algorithms and software libraries, and the supporting runtime capabilities to achieve scalable performance and resilience on heterogeneous architectures. The focus is on a critical set of fundamental linear algebra operations including direct and iterative solvers for dense and sparse linear systems of equations and eigenvalue problems. Achieving this requires a co-design effort due to the characteristics and overwhelming complexity and immense scale of such systems. Recognized experts in algorithm design and theory, parallelism, and auto-tuning will work together to explore and negotiate the necessary tradeoffs. The main research objectives are: (i) development of novel algorithms that expose as much parallelism as possible, exploit heterogeneity, avoid communication bottlenecks, respond to escalating fault rates, and help meet emerging power constraints; (ii) exploration of advanced scheduling strategies and runtime systems focusing on the extreme scale and strong scalability in multi/many-core and hybrid environments; (iii) design and evaluation of novel strategies and software support for both offline and online auto-tuning. The validation and dissemination of results will be done by integrating new software solutions into challenging scientific applications in materials science, power systems, study of energy solutions, and data analysis in astrophysics. The deliverables also include a sustainable set of methods and tools for cross-cutting issues such as scheduling, auto-tuning, and algorithm-based fault tolerance packaged into open-source library modules.

INTERTWINE - Programming Model INTERoperability ToWards Exascale (INTERTWinE)

This project addresses the problem of programming model design and implementation for the Exascale. The first Exascale computers will be very highly parallel systems, consisting of a hierarchy of architectural levels. To program such systems effectively and portably, programming APIs with efficient and robust implementations must be ready in the appropriate timescale. A single, “silver bullet” API which addresses all the architectural levels does not exist and seems very unlikely to emerge soon enough. We must therefore expect that using combinations of different APIs at different system levels will be the only practical solution in the short to medium term. Although there remains room for improvement in individual programming models and their implementations, the main challenges lie in interoperability between APIs. It is this interoperability, both at the specification level and at the implementation level, which this project seeks to address and to further the state of the art. INTERTWinE brings together the principal European organisations driving the evolution of programming models and their implementations. The project will focus on seven key programming APIs: MPI, GASPI, OpenMP, OmpSs, StarPU, QUARK and PaRSEC, each of which has a project partner with extensive experience in API design and implementation. Interoperability requirements, and evaluation of implementations will be driven by a set of kernels and applications, each of which has a project partner with a major role in their development. The project will implement a co- design cycle, by feeding back advances in API design and implementation into the applications and kernels, thereby driving new requirements and hence further advances.

greenFLASH - reen Flash, energy efficient high performance computing for real-time science

The main goal of Green Flash is to design and build a prototype for a Real-Time Controller (RTC) targeting the European Extremely Large Telescope (E-ELT) Adaptive Optics (AO) instrumentation. The E-ELT is a 39m diameter telescope to see first light in the early 2020s. To build this critical component of the telescope operations, the astronomical community is facing technical challenges, emerging from the combination of high data transfer bandwidth, low latency and high throughput requirements, similar to the identified critical barriers on the road to Exascale. With Green Flash, we will propose technical solutions, assess these enabling technologies through prototyping and assemble a full scale demonstrator to be validated with a simulator and tested on sky. With this R&D program we aim at feeding the E-ELT AO systems preliminary design studies, led by the selected first-light instruments consortia, with technological validations supporting the designs of their RTC modules. Our strategy is based on a strong interaction between academic and industrial partners. Components specifications and system requirements are derived from the AO application. Industrial partners lead the development of enabling technologies aiming at innovative tailored solutions with potential wide application range. The academic partners provide the missing links in the ecosystem, targeting their application with mainstream solutions. This increases both the value and market opportunities of the developed products. A prototype harboring all the features is used to assess the performance. It also provides the proof of concept for a resilient modular solution to equip a large scale European scientific facility, while containing the development cost by providing opportunities for return on investment.

READEX - Runtime Exploitation of Application Dynamism for Energy-efficient eXascale computing

High Performance Computing (HPC) has become a major instrument for many scientific and industrial fields to generate new insights and product developments. There is a continuous demand for growing compute power, leading to a constant increase in system size and complexity. Efficiently utilizing the resources provided on Exascale systems will be a challenging task, potentially causing a large amount of underutilized resources and wasted energy. Parameters for adjusting the system to application requirements exist both on the hardware and on the system software level but are mostly unused today. Moreover, accelerators and co-processors offer a significant performance improvement at the cost of increased overhead, e.g., for data-transfers. While HPC applications are usually highly compute intensive, they also exhibit a large degree of dynamic behaviour, e.g., the alternation between communication phases and compute kernels. Manually detecting and leveraging this dynamism to improve energy-efficiency is a tedious task that is commonly neglected by developers. However, using an automatic optimization approach, application dynamism can be detected at design-time and used to generate optimized system configurations. A light-weight run-time system will then detect this dynamic behaviour in production and switch parameter configurations if beneficial for the performance and energy-efficiency of the application. The READEX project will develop an integrated tool-suite and the READEX Programming Paradigm to exploit application domain knowledge, together achieving an improvement in energy-efficiency of up to 22.5%. Driven by a consortium of European experts from academia, HPC resource providers, and industry, the READEX project will develop a tools-aided methodology to exploit the dynamic behaviour of applications to achieve improved energy-efficiency and performance. The developed tool-suite will be efficient and scalable to support current and future extreme scale systems.

ALLScale - An Exascale Programming, Multi-objective Optimisation and Resilience Management Environment Based on Nested Recursive Parallelism

Extreme scale HPC systems impose significant challenges for developers aiming at obtaining applications efficiently utilising all available resources. In particular, the development of such applications is accompanied by the complex and labour-intensive task of managing parallel control flows, data dependencies and underlying hardware resources – each of these obligations constituting challenging problems on its own. The AllScale environment, the focus of this project, will provide a novel, sophisticated approach enabling the decoupling of the specification of parallelism from the associated management activities during program execution. Its foundation is a parallel programming model based on nested recursive parallelism, opening up the potential for a variety of compiler and runtime system based techniques adding to the capabilities of resulting applications. These include the (i) automated porting of application from small- to extreme scale architectures, (ii) the flexible tuning of the program execution to satisfy trade-offs among multiple objectives including execution time, energy and resource usage, (iii) the management of hardware resources and associated parameters (e.g. clock speed), (iv) the integration of resilience management measures to compensate for isolated hardware failures and (v) the possibility of online performance monitoring and analysis. All these services will be provided in an application independent, reusable fashion by a combination of sophisticated, modular, and customizable compiler and runtime system based solutions. AllScale will boost the development productivity, portability, and runtime, energy, and resource efficiency of parallel applications targeting small to extreme scale parallel systems by leveraging the inherent advantages of nested recursive parallelism, and will be validated with applications from fluid dynamics, environmental hazard and space weather simulations provided by SME, industry and scientific partners.

ExaFLOW - Enabling Exascale Fluid Dynamics Simulations

We are surrounded by moving fluids (gases and liquids), be it during breathing or the blood flowing in arteries; the flow around cars, ships, and airplanes; the changes in cloud formations or the plankton transport in oceans; even the formation of stars and galaxies are closely modeled as phenomena in fluid dynamics. Fluid Dynamics (FD) simulations provide a powerful tool for the analysis of such fluid flows and are an essential element of many industrial and academic problems. The complexities and nature of fluid flows, often combined with problems set in open domains, implies that the resources needed to computationally model problems of industrial and academic relevance is virtually unbounded. FD simulations therefore are a natural driver for exascale computing and have the potential for substantial societal impact, like reduced energy consumption, alternative sources of energy, improved health care, and improved climate models. The main goal of this project is to address algorithmic challenges to enable the use of accurate simulation models in exascale environments. Driven by problems of practical engineering interest we focus on important simulation aspects including: • error control and adaptive mesh refinement in complex computational domains, • resilience and fault tolerance in complex simulations • heterogeneous modeling • evaluation of energy efficiency in solver design • parallel input/output and in-situ compression for extreme data. The algorithms developed by the project will be prototyped in major open-source simulation packages in a codesign fashion, exploiting software engineering techniques for exascale. We are building directly on the results of previous exascale projects (CRESTA, EPiGRAM, etc.) and will exploit advanced and novel parallelism features required for emerging exascale architectures. The results will be validated in a number of pilot applications of concrete practical importance in close collaboration with industrial partners.

ANTAREX - AutoTuning and Adaptivity appRoach for Energy efficient eXascale HPC systems

Energy-efficient heterogeneous supercomputing architectures need to be coupled with a radically new software stack capable of exploiting the benefits offered by the heterogeneity at all the different levels (supercomputer, job, node) to meet the scalability and energy efficiency required by Exascale supercomputers. ANTAREX will solve these challenging problems by proposing a disruptive holistic approach spanning all the decision layers composing the supercomputer software stack and exploiting effectively the full system capabilities (including heterogeneity and energy management). The main goal of the ANTAREX project is to provide a breakthrough approach to express application self-adaptivity at design-time and to runtime manage and autotune applications for green and heterogenous High Performance Computing (HPC) systems up to the Exascale level.

ExaHyPE - An Exascale Hyperbolic PDE Engine

Many aspects of our life, but also cutting-edge research questions, hinge on the solution of large systems of partial differential equations expressing conservation laws. Such equations are solved to compute accurate weather forecast, complex earthquake physics, hematic flows in patients, or the most catastrophic events in the universe. Yet, our ability to exploit the predictive power of these models is still severely limited by the computational costs of their solution. Thus, the simulation of earthquakes and their induced hazards is not yet accurate enough to prevent human losses. And our ability to model astrophysical objects is still insufficient to explain our observations. While exascale supercomputers promise the performance to tackle such problems, current numerical methods are either too expensive, because not sufficiently accurate, or too inefficient, because unable to exploit the latest supercomputing hardware. Exascale software needs to be redesigned to meet the disruptive hardware changes caused by severe constraints in energy consumption. We thus develop a new exascale hyperbolic simulation engine based on high-order communication-avoiding Finite-Volume/Discontinuous-Galerkin schemes yielding high computational efficiency. We utilize structured, spacetree grids that offer dynamic adaptivity in space and time at low memory footprint. And we consequently optimise all compute kernels to minimise energy consumption and exploit inherent fault-tolerance properties of the numerical method. As a general hyperbolic solver, the exascale engine will drive research in diverse areas and relieve scientist from the burden of developing robust and efficient exascale codes. Its development is driven by precise scientific goals, addressing grand challenges in geo- and astrophysics, such as the dynamic rupture processes and subsequent regional seismic wave propagation, or the modeling of relativistic plasmas in the collision of compact stars and explosive phenomena.