E-Infrastructures
H2020-EINFRA-2016-2017

EINFRA-11-2016: Support to the next implementation phase of Pan-European High Performance Computing Infrastructure and Services (PRACE)

PRACE-5IP

PRACE Fifth Implementation Phase Project

Grant Agreement Number: EINFRA-730913

D5.1

Market and Technology Watch Report Year 1
Final

Version: 0.54
Author(s): Nico SANNA (CINECA), Aris SOTIROPOULOS (GRNET)
Date: 30.04.2018

Project and Deliverable Information Sheet

<table>
<tr><td rowspan="6">PRACE Project</td><td colspan="2">Project Ref. №: EINFRA-730913</td></tr>
<tr><td colspan="2">Project Title: PRACE Fifth Implementation Phase Project</td></tr>
<tr><td colspan="2">Project Web Site: http://www.prace-ri.eu</td></tr>
<tr><td colspan="2">Deliverable ID: D5.1</td></tr>
<tr><td colspan="2">Deliverable Nature: Report</td></tr>
<tr><td rowspan="2">Dissemination Level:
PU*</td><td>Contractual Date of Delivery:
30 / April / 2018</td></tr>
<tr><td>Actual Date of Delivery:
30 / April / 2018</td></tr>
<tr><td colspan="2">EC Project Officer: Leonardo Flores Añover</td></tr>
</table>

* - The dissemination levels are indicated as follows: **PU** – Public, **CO** – Confidential, only for members of the consortium (including the Commission Services) **CL** – Classified, as referred to in Commission Decision 2005/444/EC.

Document Control Sheet

Document	**Title:**	Market and Technology Watch Report Year 1	
	ID:	D5.1	
	Version:	0.54	**Status:** ***Final***
	Available at:	http://www.prace-ri.eu	
	Software Tool:	Microsoft Word (Windows and Mac)	
	File(s):	PRACE-5IP-D5.1.docx	
Authorship	**Written by:**		Nico SANNA (CINECA), Aris SOTIROPOULOS (GRNET)
	Contributors:		Adem Tekin (UHEM) Ahmet Tuncer Durak (UHEM) Andreas Johansson (LiU) Andrew Emerson (CINECA) Aristeidis Sotiropoulos (GRNET) Carlo Cavazzoni (CINECA) Daniele Ottaviani (CINECA) Dirk Pleiter (JSC-GCS) Eric Boyer (GENCI) Evangelia Athanasaki (GRNET) Federico Ficarelli (CINECA) Gert Svensson (KTH) Filip Stanek (IT4I-VSB) François Robin (CEA) Guillame Colin de Verdière (CEA) Javier Bartolomé (BSC) Jean-Philippe Nominé (CEA) Krzysztof Wadówka (PSNC) Marcel Bruckner (BSC) Mateusz Tykierko (WNSC) Nico Sanna (CINECA) Norbert Meyer (PNSC) Panayiotis Tsanakas (GRNET) Radek Januszewski (PSNC) Samuli Saarinen (CSC) Sena Efsun Cebeci (UHEM) Susanna Salminen (CSC)
	Reviewed by:		Stelios Erotokritou (CaSToRC) Florian Berberich (Juelich)
	Approved by:		MB/TB

Document Status Sheet

Version	Date	Status	Comments
0.10	20/February/2018	Draft	Aris Sotiropoulos (GRNET) – Initial TOC
0.12	05/March/2018	Draft	Contribution Guillaume Colin de Verdière (CEA) – 5.4
0.20	13/March/2018	Draft	Contribution Adem Tekin, Sena Efsun Cebeci Ahmet Tuncer Durak (UHEM) – 1.1
0.21	17/March/2018	Draft	Bibliography
0.22	17/March/2018	Draft	Susanna Salminen (CSC) – Ch. 3, ISC2017 vendors
0.23	17/March/2018	Draft	Adem Tekin (UHEM) Marcel Bruckner (BSC) – 2.1, INTEL, AMD, POWER
0.24	19/March/2018	Draft	Filip Stanek (IT4I) – Ch. 3. Additions to ATOS, Cray and HPE
0.25	21/March/2018	Draft	Aris Sotiropoulos (GRNET) – 1.3 Business Analysis
0.26	22/March/2018	Draft	Samuli Saarinen (CSC) – 1.5 Cloud computing and virtualization
0.27	22/March/2018	Draft	Filip Stanek (IT4I) – Ch. 2.4 Interconnects
0.28	22/March/2018	Draft	Filip Stanek (IT4I) – Revised version with the chapters 2.4.3 and 2.4.4
0.30	22/March/2018	Draft	Eric Boyer (GENCI) Proof read add-ons
0.31	23/March/2018	Draft	Eric Boyer (GENCI) Nec VE details
0.32	23/March/2018	Draft	Eric Boyer (GENCI)
0.34	25/March/2018	Draft	Carlo Cavazzoni (CINECA), Federico Ficarelli (CINECA) – 2.2 Highly parallel components/ compute engines
0.35	25/March/2018	Draft	Andreas Johansson (LiU) – 2.3.3 Tapes
0.36	25/March/2018	Draft	Norbert Meyer (PSNC) – Chapter 4
0.37	28/March/2018	Draft	Gert Svensson (KTH) – Multiple corrections
0.38	28/March/2018	Draft	Carlo Cavazzoni (CINECA), Daniele Ottaviani (CINECA) – 5.2 Quantum Computing
0.39	28/March/2018	Draft	Norbert Meyer (PNSC) – Ch. 4 update
0.40	28/March/2018	Draft	Filip Stanek (IT4I) – 2.4.3 Numalink, GenZ amendment
0.41	28/March/2018	Draft	Jean-Philippe Nominé (CEA) – 1.4 EU HPC projects landscape
0.42	29/March/2018	Draft	Eric Boyer (GENCI) – 1.4 3IP & HBP PCPs parts
0.43	30/March/2018	Draft	Mateusz Tykierko (WNSC) – 1.6 Consolidation in HPC market
0.44	30/March/2018	Draft	Nico Sanna (CINECA) – Executive Summary and Ch 1 Introduction (Ch numbers changed), Carlo Cavazzoni (CINECA), Federico Ficarelli (CINECA) – 3.3.2. NVM
0.45	30/March/2018	Prefinal	Eric Boyer (GENCI) – Updated NEC section, Aris Sotiropoulos (GRNET) var corrections and additions towards prefinal version
0.46	30/March/2018	Prefinal	Dirk Pleiter – 6.1 Artificial intelligence & Deep Learning, 6.3 Neuromorphic computing
0.47	03/April/2018	Prefinal	Jean-Philippe Nominé (CEA) – 2.1 Update on exascale initiatives: China, Japan, USA and Europe
0.48	03/April/2018	Prefinal	Filip Stanek (IT4I) – 3.3 Memory and storage technologies, 3.3.1 HBM, HMC and GDDR, 3.3.2 DRAM
0.49	04/April/2018	Prefinal	François Robin (CEA) – var corrections
0.50	04/April/2018	Prefinal	Jean-Philippe Nominé (CEA) – Add PPI4HPC to 2.4
0.51	10/April/2018	Internal Review	Nico Sanna (CINECA) – Executive Summary updated
0.52	17/April/2018	After Int. Rev.	Aris Sotiropoulos (GRNET)
0.52a	23/April/2018	Prefinal	Nico Sanna (CINECA) – Annotated corrected version w/ revisions
0.53	23/April/2018	Final version	Nico Sanna (CINECA)
0.54	30/April/2018	Final version	Nico Sanna (CINECA) – Corrected version after MB/TB review

Document Keywords

Keywords:	PRACE, HPC, Research Infrastructure, Market Watch, Technology, Exascale, TOP500, Horizon 2020, Green 500, Benchmarks, Computing Efficiency, Energy efficiency, Cloud computing, Virtualization, CPU, Data Storage, Data Services

Table of Contents

List of Figures

List of Tables

References and Applicable Documents

[1] "TOP500," 2017. [Online]. Available: www.top500.org.

[2] "Green500 Web page," 2017. [Online]. Available: www.top500.org/green500/.

[3] "High Performance Conjugate Gradients," [Online]. Available: www.hpcg-benchmark.org.

[4] "PRACE-4IP D5.2 Deliverable "Market and Technology Watch Report Year 2. Final summary of results gathered"," 2017.

[5] "PRACE-4IP D5.1 Deliverable "Market and Technology Watch Report Year 1"," 2016.

[6] T. N. Team, "China Pulls Ahead of U.S. in Latest TOP500 List," 13 11 2017. [Online]. Available: https://www.top500.org/news/china-pulls-ahead-of-us-in-latest-top500-list/. [Accessed 3 4 2018].

[7] [Online]. Available: http://www.r-ccs.riken.jp/en/postk/project .

[8] "Tsubame 3, Japan's 'AI' supercomputer became operational 1st August 2017," [Online]. Available: https://www.nextplatform.com/2017/08/22/inside-view-tokyo-techs-massive-tsubame-3-supercomputer/.

[9] [Online]. Available: https://www.nitrd.gov/nsci/ .

[10] [Online]. Available: https://www.olcf.ornl.gov/olcf-resources/compute-systems/summit/ .

[11] [Online]. Available: https://computation.llnl.gov/computers/sierra .

[12] [Online]. Available: https://www.hpcwire.com/2017/09/27/us-coalesces-plans-first-exascale-supercomputer-aurora-2021/ .

[13] [Online]. Available: https://www.exascaleproject.org/ .

[14] [Online]. Available: https://ec.europa.eu/digital-single-market/high-performance-computing-hpc.

[15] [Online]. Available: https://ec.europa.eu/digital-single-market/en/european-cloud-initiative.

[16] [Online]. Available: https://ec.europa.eu/digital-single-market/en/high-performance-computing-contractual-public-private-partnership-hpc-cppp.

[17] Timothy Prickett Morgan, "Casing The HPC Market Is Hard, And Getting Harder," 22 June 2017. [Online]. Available: https://www.nextplatform.com/2017/06/22/casing-hpc-market-hard-getting-harder/. [Accessed 19 March 2018].

[18] [Online]. Available: https://ec.europa.eu/digital-single-market/en/high-performance-computing .

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

[20] "ETP4HPC," [Online]. Available: www.etp4hpc.eu.

[21] [Online]. Available: https://ec.europa.eu/programmes/horizon2020/en/news/overview-eu-funded-centres-excellence-computing-applications.

[22] [Online]. Available: https://ec.europa.eu/digital-agenda/en/high-performance-computing-contractual-public-private-partnership-hpc-cppp .

[23] [Online]. Available: http://ec.europa.eu/programmes/horizon2020/en/h2020-section/leadership-enabling-and-industrial-technologies.

[24] [Online]. Available: http://www.etp4hpc.eu/sra.html.

[25] [Online]. Available: http://www.etp4hpc.eu/european-hpc-handbook.html.

[26] "Public Procurement of Innovations for High Performance Computing," [Online]. Available: https://www.ppi4hpc.eu/. [Accessed 04 04 2018].

[27] [Online]. Available: https://cordis.europa.eu/project/rcn/209998_en.html.

[28] [Online]. Available: https://newsroom.intel.com/editorials/intels-stratix-10-fpga-supporting-smart-connected-revolution/.

[29] [Online]. Available: https://www.altera.com/en_US/pdfs/literature/wp/wp-01220-hyperflex-architecture-fpga-socs.pdf.

[30] [Online]. Available: https://www.altera.com/content/dam/altera-www/global/en_US/pdfs/literature/wp/wp-01231-understanding-how-hyperflex-architecture-enables-high-performance-systems.pdf.

[31] [Online]. Available: https://www.altera.com/content/dam/altera-www/global/en_US/pdfs/literature/backgrounder/stratix10-floating-point-backgrounder.pdf.

[32] [Online]. Available: https://www.altera.com/content/dam/altera-www/global/en_US/pdfs/literature/po/ps-opencl.pdf.

[33] [Online]. Available: https://fuse.wikichip.org/news/191/the-2048-core-pezy-sc2-sets-a-green500-record/.

[34] [Online]. Available: https://pdfs.semanticscholar.org/25ec/4af90d4c02510b0b589d50fa01a3c0536e72.pdf.

[35] [Online]. Available: https://en.wikichip.org/wiki/pezy/pezy-scx.

[36] [Online]. Available: https://www.hpcwire.com/2017/03/14/new-japanese-supercomputing-project-targets-exascale.

[37] [Online]. Available: https://riscv.org/contributors/.

[38] [Online]. Available: https://riscv.org/specifications/.

[39] [Online]. Available: https://www.pulp-platform.org/publications/.

[40] [Online]. Available: http://primeurmagazine.com/flash/AE-PR-03-18-87.html.

[41] [Online]. Available: http://www.adapteva.com/andreas-blog/why-i-will-be-using-the-risc-v-in-my-next-chip.

[42] [Online]. Available: https://www.youtube.com/watch?v=gg1lISJfJI0.

[43] [Online]. Available: https://riscv.org/2017/11/ee-times-article-risc-v-spins-drives-ai/.

[44] [Online]. Available: https://riscv.org/2018/03/designnews-article-first-open-source-risc-v-soc-linux-released/.

[45] [Online]. Available: https://images.nvidia.com/content/technologies/volta/pdf/tesla-volta-v100-datasheet-letter-fnl-web.pdf.

[46] [Online]. Available: https://www-01.ibm.com/common/ssi/cgi-bin/ssialias?subtype=SP&infotype=PM&appname=STGE_DC_ZQ_USEN&htmlfid=DCD12374USEN.

[47] [Online]. Available: https://arxiv.org/pdf/1704.08273.pdf.

[48] [Online]. Available: https://www.anandtech.com/show/12301/samsung-starts-production-of-hbm2-aquabolt-memory-8-gb-24-gbps.

[49] [Online]. Available: https://www.gamersnexus.net/industry/3271-gddr6-slated-for-next-gen-nvidia-gpus-3-month-mass-production.

[50] [Online]. Available: http://www.snia.org/sites/default/files/DSI/2016/presentations/stor_class_mem/SarahJelinek__Preparing_for_PersistentMemory16.pdf.

[51] [Online]. Available: https://software.intel.com/en-us/persistent-memory.

[52] [Online]. Available: https://www.snia.org/sites/default/files/PM-Summit/2018/presentations/03_PMSummit_18_Rudoff_Final_Post.pdf.

[53] M. S. Woodacre, *HPE, Meeting at SC17*, Denver, CO, US, 2017.

[54] "PRACE-4IP Deliverable 5.1," [Online]. Available: http://www.prace-ri.eu/IMG/pdf/D5.1_4ip.pdf.

[55] "Express The Promise and Progress of NVM," [Online]. Available: https://www.hpcwire.com/2015/10/29/hpc-eyes-non-volatile-memory-express/.

[56] [Online]. Available: http://www.brightcomputing.com/blog/taking-the-roadblocks-out-of-hpc-with-nvme.

[57] "NVMe 6 Reasons to Consider.," [Online]. Available: https://www.siliconmechanics.com/i63334/6-reasons-to-consider-nvme.php.

[58] H. d. s. t. t. f. 2017, http://searchstorage.techtarget.com/feature/Hot-data-storage-technology-trends-for-2017.

[59] "Intel® Omni-Path Fabric 100 Series," [Online]. Available: https://www.intel.com/content/www/us/en/high-performance-computing-fabrics/omni-path-architecture-fabric-overview.html.

[60] "Exploring Intel's Omni-Path Network Fabric.," [Online]. Available: http://www.anandtech.com/show/9561/exploring-intels-omnipath-network-fabric.

[61] "Exascaler," [Online]. Available: http://www.ddn.com/products/lustre-file-system-exascaler..

[62] "Gridscaler," [Online]. Available: http://www.ddn.com/products/parallel-file-system-gridscaler.

[63] [Online]. Available: http://www.computerweekly.com/news/2240231962/DDN-launches-scale-out-NAS-GS7K-GPFS-product-bundle.

[64] [Online]. Available: https://www.beegfs.io/docs/whitepapers/Introduction_to_BeeGFS_by_ThinkParQ.pdf.

[65] Fujitsu, "Post-K Incorporating ARM SVE: Power for Emerging Apps," [Online]. Available: http://www.fujitsu.com/global/Images/post-k-incorporating-arm-sve-power-for-emerging-apps.pdf. [Accessed 30 March 2018].

[66] IBM, "IBM Power System AC922 Introduction and Technical Overview," [Online]. Available: https://www.redbooks.ibm.com/Redbooks.nsf/RedpieceAbstracts/redp5472.html. [Accessed 30 March 2018].

[67] AMD, "Radeon Instrinct M25," [Online]. Available: https://instinct.radeon.com/en/product/mi/radeon-instinct-mi25/. [Accessed 30 March 2018].

[68] NVIDIA, "NVIDIA TESLA V100 GPU Architecture," [Online]. Available: http://images.nvidia.com/content/volta-architecture/pdf/volta-architecture-whitepaper.pdf. [Accessed 30 March 2018].

[69] P. Alcorn, "Hot Chips 2017: A Closer Look At Google's TPU v2," [Online]. Available: http://www.tomshardware.com/news/tpu-v2-google-machine-learning,35370.html. [Accessed 30 March 2018].

[70] Fujitsu, "Post-K Development and Introducing DLU," [Online]. Available: http://www.fujitsu.com/global/Images/post-k-development-and-introducing-dlu.pdf. [Accessed 30 March 2018].

[71] Xilinx, "Machine Learning Inference Solutions from Edge to Cloud," [Online]. Available: https://www.xilinx.com/applications/megatrends/machine-learning.html. [Accessed 30 March 2018].

[72] C. B. A. C. P. Schaar, "Cognitive Computing for Materials Design: A Metallurgy Case Study," [Online]. Available: https://www.zurich.ibm.com/pdf/science-posters/5_Staar_MaterialsDesign_Poster_Aug20_2015.pdf. [Accessed 30 March 2018].

[73] ORNL, "Exascale Deep Learning and Simulation Enabled Precision Medicine for Cancer," [Online]. Available: http://candle.cels.anl.gov/. [Accessed 30 March 2018].

[74] J. HPC, "JADE system specification," [Online]. Available: http://www.jade.ac.uk/. [Accessed 30 March 2018].

[75] S. Matsuoka, "TSUBAME3 and ABCI: Supercomputer Architectures for HPC and AI / BD Convergence," [Online]. Available: http://on-demand.gputechconf.com/gtc/2017/presentation/S7813-Matsuoka-scalable.pdf.pdf. [Accessed 30 March 2018].

[76] ORNL, "Summit: Oak Ridge National Laboratory's next High Performance Supercomputer," [Online]. Available: https://www.olcf.ornl.gov/olcf-resources/compute-systems/summit/ . [Accessed 30 March 2018].

[77] R. Smith, "NVIDIA Develops NVLink Switch: NVSwitch, 18 Ports for DGX-2 & More," [Online]. Available: https://www.anandtech.com/show/12581/nvidia-develops-nvlink-switch-nvswitch-18-ports-for-dgx2-more. [Accessed 30 March 2018].

[78] A. Sergeev and M. Del Balso, "Horovod: fast and easy distributed deep learning in TensorFlow," arXiv, Ithaca, 2018.

[79] P. Messina and S. Lee, "Exascale Computing Project Update," [Online]. Available: https://science.energy.gov/~/media/ascr/ascac/pdf/meetings/201709/Paul_Messina_ECP_Update_ASCAC__20170927.pdf.

[80] Y. C. Y.-H. L. S.-K. L. L. Z. J.-G. R. W.-Q. C. W.-Y. L. Juan Yin, "Satellite-based entanglement distribution over 1200 kilometers," *Science*, vol. 356, no. 6343, pp. 1140-1144, 2017.

[81] [Online]. Available: https://en.wikipedia.org/wiki/List_of_companies_involved_in_quantum_computing_or_communication.

[82] [Online]. Available: http://www.scmp.com/tech/china-tech/article/2134520/alibaba-cloud-steps-its-game-it-offers-quantum-computing-service.

[83] [Online]. Available: https://quantumexperience.ng.bluemix.net/qx/experience.

[84] J. A. G. G. N. L. H. T. M. E.-g. S. a. R. W. Edwin Pednault, "Breaking the 49-Qubit Barrier in the Simulation of Quan-tum Circuits," 16 October 2017. [Online]. Available: arXiv:1710.05867v1 [quant-ph].

[85] A. M. K. T. Abhinav Kandala, "Hardware-efficient Variational Quantum Eigensolver for Small Molecules and Quantum Magnets," 13 October 2017. [Online]. Available: arXiv:1704.05018v2 [quant-ph].

[86] J. Schemmel and e. al., "A wafer-scale neuromorphic hardware system for large-scale neural modeling," ISCAS 2010 Proceedings, 2010.

[87] "NICE 2018 workshop," [Online]. Available: http://niceworkshop.org/2018-nice-workshop/.

[88] H. B. Project, "Neuromorphic Computing Platform," [Online]. Available: https://www.humanbrainproject.eu/ncp.

[89] T. Sharp, C. Patterson and S. Furber, "Distributed Configuration of Massively-Parallel Simulation on SpiNNaker Neuromorphic Hardware," International Joint Conference on Neural Networks, X, 2011.

[90] Intel, "Intel’s New Self-Learning Chip Promises to Accelerate Artificial Intelligence," 25 September 2017. [Online]. Available: https://newsroom.intel.com/editorials/intels-new-self-learning-chip-promises-accelerate-artificial-intelligence/. [Accessed 30 March 2018].

[91] J. Sawada and e. al., "TrueNorth Ecosystem for Brain-Inspired Computing: Scalable Systems, Software, and Applications," SC16 Proceedings, 2016.

[92] IBM, "U.S. Air Force Research Lab Taps IBM to Build Brain-Inspired AI Supercomputing System," 23 June 2017. [Online]. Available: https://www-03.ibm.com/press/us/en/pressrelease/52657.wss. [Accessed 30 March 2018].

[93] "Trinity at LANL," [Online]. Available: http://www.lanl.gov/projects/trinity/.

[94] "Tera 1000 at CEA," [Online]. Available: http://www-hpc.cea.fr/en/complexe/tera.htm.

[95] "Knights Landing KNL," [Online]. Available: https://ark.intel.com/products/codename/48999/Knights-Landing.

[96] "EuroExa project," [Online]. Available: https://euroexa.eu/.

[97] "Intel Nervana Neural Network Processors (NNP) Redefine AI Silicon," [Online]. Available: https://ai.intel.com/intel-nervana-neural-network-processors-nnp-redefine-ai-silicon/.

[98] "Shor's algorithm," [Online]. Available: https://en.wikipedia.org/wiki/Shor%27s_algorithm.

[99] "PRACE Web site," [Online]. Available: www.prace-ri.eu.

[100] [Online]. Available: https://www.altera.co.jp/content/dam/altera-www/global/en_US/pdfs/literature/wp/wp-01253-leveraging-stratix10-hyperflex-for-maximum-power.pdf.

[101] [Online]. Available: https://en.wikichip.org/wiki/zettascaler.

[102] [Online]. Available: http://images.nvidia.com/content/volta-architecture/pdf/volta-architecture-whitepaper.pdf.

[103] T. P. a. P. o. N. Express, https://www.hpcwire.com/2015/10/29/hpc-eyes-non-volatile-memory-express/.

[104] http://www.brightcomputing.com/blog/taking-the-roadblocks-out-of-hpc-with-nvme.

[105] 6. R. t. C. NVMe, https://www.siliconmechanics.com/i63334/6-reasons-to-consider-nvme.php.

[106] I. O.-P. F. 1. S. -, https://www.intel.com/content/www/us/en/high-performance-computing-fabrics/omni-path-architecture-fabric-overview.html.

[107] E. I. O.-P. N. F. -, http://www.anandtech.com/show/9561/exploring-intels-omnipath-network-fabric.

[108] E. -. [Online]. Available: http://www.ddn.com/products/lustre-file-system-exascaler.

[109] G. -. [Online]. Available: http://www.ddn.com/products/parallel-file-system-gridscaler.

[110] [Online]. Available: http://www.computerweekly.com/news/2240231962/DDN-launches-scale-out-NAS-GS7K-GPFS-product-bundle.

[111] [Online]. Available: https://www.beegfs.io/docs/whitepapers/Introduction_to_BeeGFS_by_ThinkParQ.pdf.

[112] [Online]. Available: https://www-01.ibm.com/common/ssi/cgi-bin/ssialias?subtype=SP&infotype=PM&appname=STGE_DC_ZQ_USEN&htmlfid=DCD12374USEN.

[113] "Knights Landing Processor with Omni-Path Makes Cloud Debut," [Online]. Available: https://www.hpcwire.com/2017/04/18/knights-landing-processor-omni-path-makes-cloud-debut/.

[114] "Hot data storage technology trends for 2017," [Online]. Available: http://searchstorage.techtarget.com/feature/Hot-data-storage-technology-trends-for-2017.

[115] [Online]. Available: https://www.technologyreview.com/s/609451/ibm-raises-the-bar-with-a-50-qubit-quantum-computer/.

[116] Y. L. Z.-q. Y. B. Z. Yuanhao Wang, "16-qubit IBM universal quantum comput-er can be fully entangled," 11 January 2018. [Online]. Available: https://arxiv.org/abs/1801.03782v1.

[117] [Online]. Available: https://events.static.linuxfound.org/sites/events/files/slides/LinuxCon_16_PersistentMemoryInLinux_0.pdf.

[118] [Online]. Available: https://ec.europa.eu/digital-single-market/en/eurohpc-joint-undertaking.

[119] [Online]. Available: https://pdfs.semanticscholar.org/0fb6/bf85471e8049269a15cbb53eb89cfc8603ee.pdf.

[120] [Online]. Available: http://www.gauss-centre.eu/SharedDocs/Meldungen/GAUSS-CENTRE/EN/2017/news_10_BMBF_smart_scale.html?nn=1282668.

[121] [Online]. Available: https://www.entreprises.gouv.fr/politique-et-enjeux/plan-supercalculateurs.

[122] [Online]. http://www.exascale.org/bdec/sites/www.exascale.org.bdec/files/China-Overview16to9_Fu.pdf

[123] Design and Implementation for Convergence of HPC and Bigdata on Tianhe2, Yutong Lu, NUDT, China - ORAP Forum 39 (http://orap.irisa.fr/?page_id=696)

List of Acronyms and Abbreviations

aisbl	Association International Sans But Lucratif (legal form of the PRACE-RI)
AI	Artificial Intelligence
ASIC	Application Specific Integrated Circuit
BXI	Bull eXascale Interconnect (product by ATOS)
CEF	Connecting Europe Facility
CoE	Center of Excellence
cPPP	contractual Public Private Partnership
CPU	Central Processing Unit
CSA	Coordination and support action (type of H2020 project)
CUDA	Compute Unified Device Architecture (NVIDIA)
DoE	(US) Department of Energy
EC	European Commission
ECP	Exascale Computing Project
ECI	European Cloud Initiative
EDI	European Data Infrastructure
EFlops	Exa (= 1018) Floating-point operations (usually in 64-bit, i.e. DP) per second, also EF/s or EF
EOSC	European Open Science Cloud
GB	Giga (= 230 ~ 109) Bytes (= 8 bits), also GByte
Gb/s	Giga (= 109) bits per second, also Gbit/s
GB/s	Giga (= 109) Bytes (= 8 bits) per second, also GByte/s
GÉANT	Collaboration between National Research and Education Networks to build a multi-gigabit pan-European network. The current EC-funded project as of 2015 is GN4.
GFlop/s	Giga (= 109) Floating point operations (usually in 64-bit, i.e. DP) per second, also GF/s
GHz	Giga (= 109) Hertz, frequency =109 periods or clock cycles per second
GPU	Graphic Processing Unit
GT/s	Giga (109) transfers per second
HDR	Infiniband interconnect generation called High Data Rate with link speed 200 Gb/s
HPC	High-Performance Computing; Computing at a high- performance level at any given time; often used synonym with Supercomputing
HPCG	High Performance Conjugate Gradients
HPL	High-Performance LINPACK
IoT	Internet of Things.
ISC	International Supercomputing Conference; European equivalent to the US based SCxx conference. Held annually in Germany.
JU	Joint Undertaking
KB	Kilo (= 210 ~103) Bytes (= 8 bits), also Kbyte
LFF	Large Form Factor
LINPACK	Software library for Linear Algebra
LTO	Linear Tape-Open
MB	Management Board (highest decision making body of the project)
MB	Mega (= 220 ~ 106) Bytes (= 8 bits), also MByte
MB/s	Mega (= 106) Bytes (= 8 bits) per second, also MByte/s

MFlop/s	Mega (= 106) Floating point operations (usually in 64-bit, i.e. DP) per second, also MF/s
MPI	Message Passing Interface
NCSI	National Strategic Computing Initiative
NDA	Non-Disclosure Agreement. Typically signed between vendors and customers working together on products prior to their general availability or announcement.
NDR	Next generation of Infiniband interconnect (after HDR)
NIC	Network Interface Controller
NRZ	Non Return to Zero binary signalling for serial links
NVM	Non Volatile Memory
PAM-4	Pulse Amplitude Modulation signalling for high-speed serial links
PFlops	Peta (= 1015) Floating-point operations (usually in 64-bit, i.e. DP) per second, also PF/s or PF
PMDK	Persistent Memory Developer Kit
PRACE	Partnership for Advanced Computing in Europe; Project Acronym
PRACE 2	The next phase of the PRACE Research Infrastructure following the initial five-year period.
QUBO	Quadratic unconstrained binary optimization
RI	Research Infrastructure
RIA	Research and innovation action (type of H2020 project)
SERDES	Serializer/Deserializer, converts data from/to serial/parallel interfaces
TB	Technical Board (group of Work Package leaders)
TB	Tera (= 240 ~ 1012) 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.
TFlops	Tera (= 1012) Floating-point operations (usually in 64-bit, i.e. DP) per second, also TF/s or TF
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
TOR	Top Of Rack, usually network/Infiniband switch connecting devices in one rack
UPI	Ultra Path Interconnect, Intel technology to link multiple CPUs in coherent way

List of Project Partner Acronyms

BADW-LRZ	Leibniz-Rechenzentrum der Bayerischen Akademie der Wissenschaften, Germany (3rd Party to GCS)
BILKENT	Bilkent University, Turkey (3rd Party to UYBHM)
BSC	Barcelona Supercomputing Center - Centro Nacional de Supercomputacion, Spain
CaSToRC	Computation-based Science and Technology Research Center, Cyprus
CCSAS	Computing Centre of the Slovak Academy of Sciences, Slovakia
CEA	Commissariat à l'Energie Atomique et aux Energies Alternatives, France (3rd Party to GENCI)
CESGA	Fundacion Publica Gallega Centro Tecnológico de Supercomputación de Galicia, Spain, (3rd Party to BSC)
CINECA	CINECA Consorzio Interuniversitario, Italy
CINES	Centre Informatique National de l'Enseignement Supérieur, France (3rd Party to GENCI)
CNRS	Centre National de la Recherche Scientifique, France (3rd Party to GENCI)
CSC	CSC Scientific Computing Ltd., Finland
CSIC	Spanish Council for Scientific Research (3rd Party to BSC)
CYFRONET	Academic Computing Centre CYFRONET AGH, Poland (3rd party to PNSC)
EPCC	EPCC at The University of Edinburgh, UK
ETHZurich (CSCS)	Eidgenössische Technische Hochschule Zürich – CSCS, Switzerland
FIS	FACULTY OF INFORMATION STUDIES, Slovenia (3rd Party to ULFME)
GCS	Gauss Centre for Supercomputing e.V., Germany
GENCI	Grand Equipement National de Calcul Intensiv, France
GRNET	Greek Research and Technology Network, Greece
INRIA	Institut National de Recherche en Informatique et Automatique, France (3rd Party to GENCI)
IST	Instituto Superior Técnico, Portugal (3rd Party to UC-LCA)
IT4Innovations	IT4Innovations National supercomputing centre at VŠB-Technical University of Ostrava, Czech Republic
IUCC	INTER UNIVERSITY COMPUTATION CENTRE, Israel
JUELICH	Forschungszentrum Juelich GmbH, Germany
KIFÜ (NIIFI)	Governmental Information Technology Development Agency, Hungary
KTH	Royal Institute of Technology, Sweden (3rd Party to SNIC)
LiU	Linkoping University, Sweden (3rd Party to SNIC)
NCSA	NATIONAL CENTRE FOR SUPERCOMPUTING APPLICATIONS, Bulgaria
NTNU	The Norwegian University of Science and Technology, Norway (3rd Party to SIGMA)
NUI-Galway	National University of Ireland Galway, Ireland
PRACE	Partnership for Advanced Computing in Europe aisbl, Belgium
PSNC	Poznan Supercomputing and Networking Center, Poland
RISCSW	RISC Software GmbH
RZG	Max Planck Gesellschaft zur Förderung der Wissenschaften e.V., Germany (3 rd Party to GCS)

SIGMA2	UNINETT Sigma2 AS, Norway
SNIC	Swedish National Infrastructure for Computing (within the Swedish Science Council), Sweden
SoC	System on Chip
STFC	Science and Technology Facilities Council, UK (3rd Party to EPSRC)
SURFsara	Dutch national high-performance computing and e-Science support center, part of the SURF cooperative, Netherlands
UC-LCA	Universidade de Coimbra, Labotatório de Computação Avançada, Portugal
UCPH	Københavns Universitet, Denmark
UHEM	Istanbul Technical University, Ayazaga Campus, Turkey
UiO	University of Oslo, Norway (3rd Party to SIGMA)
ULFME	UNIVERZA V LJUBLJANI, Slovenia
UmU	Umea University, Sweden (3rd Party to SNIC)
UnivEvora	Universidade de Évora, Portugal (3rd Party to UC-LCA)
UPC	Universitat Politècnica de Catalunya, Spain (3rd Party to BSC)
UPM/CeSViMa	Madrid Supercomputing and Visualization Center, Spain (3rd Party to BSC)
USTUTT-HLRS	Universitaet Stuttgart – HLRS, Germany (3rd Party to GCS)
WCNS	Politechnika Wroclawska, Poland (3rd Party to PNSC)

Executive Summary

This document is the first deliverable of PRACE-5IP Work Package 5 “Task 5.1 - Technology and market watch” and corresponds to a periodic annual update on technology and market trends. It is thus the continuation of a well-established effort to carry out an assessment of the HPC market based on market surveys, supercomputing conferences, and exchanges between vendors and experts involved in the work package.

In summary, the TOP500 list is still dominated by systems based in China, but Japan is emerging in the Green500 list where the first EU system is ranked #10. In the last year, increased attention has been given to the HPCG benchmark list where the EU countries are ranked from #13 to #28. Overall, the number of EU systems decreased between November 2016 and November 2017 in all lists which were analysed. Among the most powerful HPC systems, NVIDIA GPUs appear to be the preferred accelerator followed by Xeon Phi, but future trends will drastically change after the announcement of its withdrawal by Intel.

Plans for exascale are well-defined in China, USA and Japan with first prototype delivery dates estimated in 2020, 2021, 2021/2022, respectively. Europe is well in line with this worldwide trend with the EuroHPC project aiming at delivering the first exascale prototype in 2022/2023.

HPC cloud services are dominated by commercial vendors while OpenStack solutions are gaining momentum - even if those technologies generally complement, rather than replace, the traditional HPC systems.

Core technologies for HPC system processors mostly rely on Intel Sky/Cascade/Ice Lake silicon development for present and near future X86 systems. AMD has emerged with the EPYC processor and IBM with its POWER9. ARM technology adoption seems to be on the rise with the release of the new SoC(s) which is comparable to the Cavium ThunderX2 with core-by-core performance similar to the Skylake processor.

Trends in accelerators for HPC systems mostly rely on NVIDIA GPUs with the Volta V100 engine, but the PEZY processor (in Japan), FPGA and RISC-V compute engines are also playing a role in the development of future HPC architectures.

3D memory subsystems are emerging as successors of current DRAM technology, while NVM DIMMS are reaching the market to support (and eventually replace) SSD in cache design. This will drastically improve I/O performance which are expected in near future high-end storage subsystems.

1 Introduction

The PRACE-5IP Work Package 5 (WP5), "HPC Commissioning and Prototyping", has three objectives:

- Technology and market watch, vendor relationships independent of procurements (Task 1);
- Best practice for energy-efficient HPC centre infrastructures design and operations (Task 2);
- Extended best practice guide for prototypes or demonstrators (Task 3).

WP5 builds on the important work performed in all previous PRACE Implementation Projects (IP) in terms of technology and market watch, know-how and best practices for energy-efficient HPC Centre infrastructures design and operations, innovative procurement of R&D and prototyping of HPC systems. It aims to deliver information and guidance useful for decision makers at different levels.

The first objective of PRACE-5IP Work Package 5 is Task 5.1 of, "Technology and market watch". It is the continuation of a well-established effort, using assessments of the HPC market based on market surveys, Top500 andGreen500/HPCG lists analyses, supercomputing conferences, and exchanges between vendors and experts who are involved in the work package. Trends and innovations based on the work of prototyping activities in previous PRACE implementation 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, Horizon 2020 FETHPC1-2014 and follow-ups in future Work Programmes.

This is the first deliverable from Task 5.1 of Work Package 5 of the PRACE-5IP project. It focusses 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 previous PRACE projects and deliverables are now dealt with in other deliverables or white papers (and tasks) of WP5.

This deliverable contains many technical details in some topics and is intended for persons who are actively working in the HPC field. Practitioners should read this document to get an overview of developments on the infrastructure side and how it may affect planning for future data centres and systems.

The deliverable is organised into five main chapters. In addition to this introduction (Chapter 1) it contains:

- Chapter 2: "Worldwide HPC landscape and market overview" which uses and analyses the TOP500, Green500 and HPCG lists with a geographical and business topical angle. It then proposes some extra considerations from other sources, as well as a brief overview of large HPC initiatives in the EU and world-wide, together with current trends on HPC cloud computing;
- Chapter 3: "Core technologies and components" is a quick overview of processors, accelerators, memory and storage technologies and interconnect technologies;
- Chapter 4: "Overview of vendor solutions" gives some vendor snapshots, and looks at some trends of technologies in the near future HPC market;

- Chapter 5: “Data storage and data management” is an overview of present storage models, architectures and solutions;
- Chapter 6: “Paradigm shifts in HPC technologies” reports on the most promising technologies for future HPC systems, including neuromorphic and quantum computing, AI and heterogeneous architectures.

2 Worldwide HPC landscape and market overview

2.1 A quick snapshot of HPC worldwide

This section provides an overview of HPC worldwide, with a special focus on Europe based on statistical data provided from the TOP500 [1], the Green 500 [2] and the HPCG [3] lists. In the subsequent analysis, special attention is given to the 10, 20 and 50 most powerful systems in the world according to the TOP500 and Green500 rankings.

2.1.1 Countries

November 2017 results for Rmax values in TOP 500 (TFlop/s), power efficiency (GFlops/watts) values in Green 500 and HPCG Benchmark (PFlop/s) are presented in Table 1.

	TOP 500	GREEN 500	HPCG
1.	China (Sunway TaihuLight)	Japan (Shoubu system B)	Japan (K Computer)
2.	China (Tianhe-2, MilkyWay 2)	Japan (Suiren2)	China (Tianhe-2, MilkyWay 2)
3.	Switzerland (Piz Daint)	Japan (Sakura)	USA (Trinity)
4.	Japan (Gyoukou)	USA (DGX SaturnV Volta)	Switzerland (Piz Daint)
5.	United States (Titan)	Japan (Gyoukou)	China (Sunway TaihuLight)
6.	United States (Sequoia)	Japan (TSUBAME3.0)	Japan (Oakforest-PACS)
7.	United States (Trinity)	Japan (AIST AI Cloud)	United States (Cori)
8.	United States (Cori)	Japan (RAIDEN GPU subsystem)	United States (Sequoia)
9.	Japan (Oakforest-PACS)	United Kingdom (Wilkes-2)	United States (Titan)
10.	Japan (K Computer)	Switzerland (Piz Daint)	Japan (TSUBAME3.0)

Table 1. Top10 systems in benchmark results for TOP500/ GREEN500/ HPCG.

The benchmarking results show that a system in the top 10 in the TOP500 list has a different rank in Green500 and HPCG based on focused parameters. For instance, Japan (Shoubu system B) takes place in the Green500, however, its rank in the TOP500 list is 259 and does not appear in the HPCG benchmark. Based on the 2017 TOP500 benchmarks, China has the first and the second fastest systems in the world. In the case of Green500 statistics, Japan apparently dominates. Japan's K computer also leads the HPCG list; here the United States is represented by four systems.
According to the TOP500 results, when compared with previous years (2016), China's attack with the Sunway TaihuLight system in the market is visible. It has an Rmax value of 93014.6 (TFlop/s) and their other system Tianhe-2 (MilkyWay-2) follows it with Rmax of 33,862.7 (TFlop/s).

The rankings for Europe in the TOP500 and HPCG benchmarks are as follows: Germany's rank in the TOP500 is 19 and in HPCG is 14; France's rank in TOP500 is 21 and in HPCG is 13**;** United Kingdom's rank in TOP500 is 15 and in HPCG is 24**;** Italy's rank in TOP500 is 14 and in HPCG is 28 and finally Spain's rank in the TOP500 is 16 and in HPCG is 15.

Figure 1 and Figure 2 present the system and performance share of the countries according to the November 2017 TOP500 benchmarks. In these figures the system share of a country is presented by the number of systems present and the system share by the total Rmax values. Since the Green500 list includes the same systems of the TOP500 in a different order, the system share results for the Green500 are identical to the TOP500.

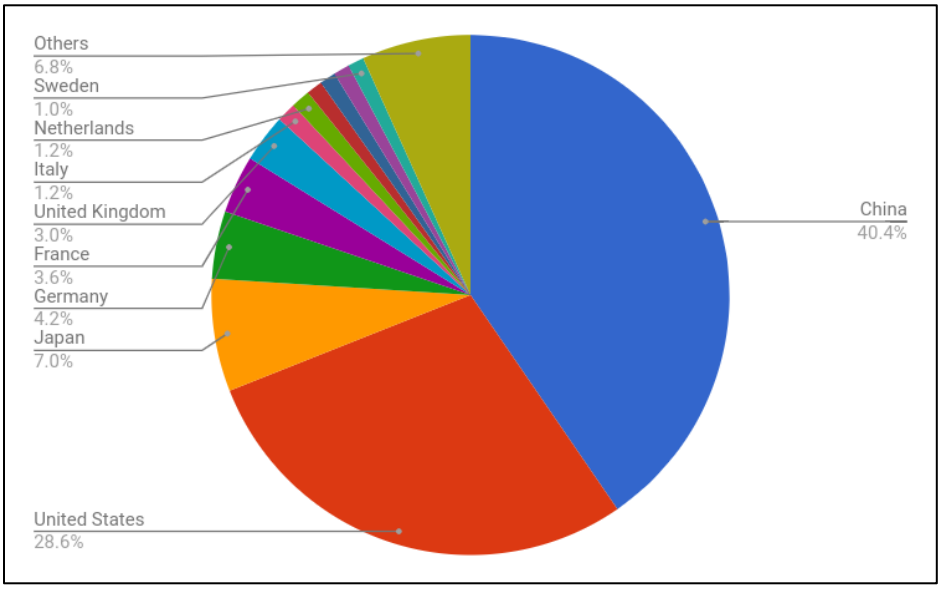

Figure 1. System share in TOP500 and Green500.

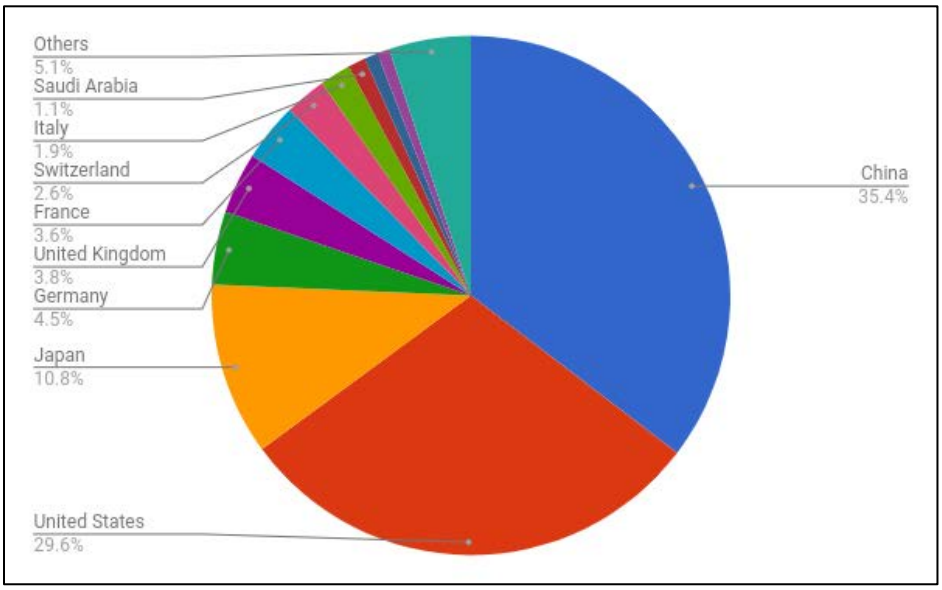

Figure 2. Performance share in TOP500.

Figure 3 shows the evolution of the presence of countries in the TOP500 list over the last five years, measured by the percentage of the number of systems. The graph shows that China is taking over the lead from the US, with most other countries also in decline.

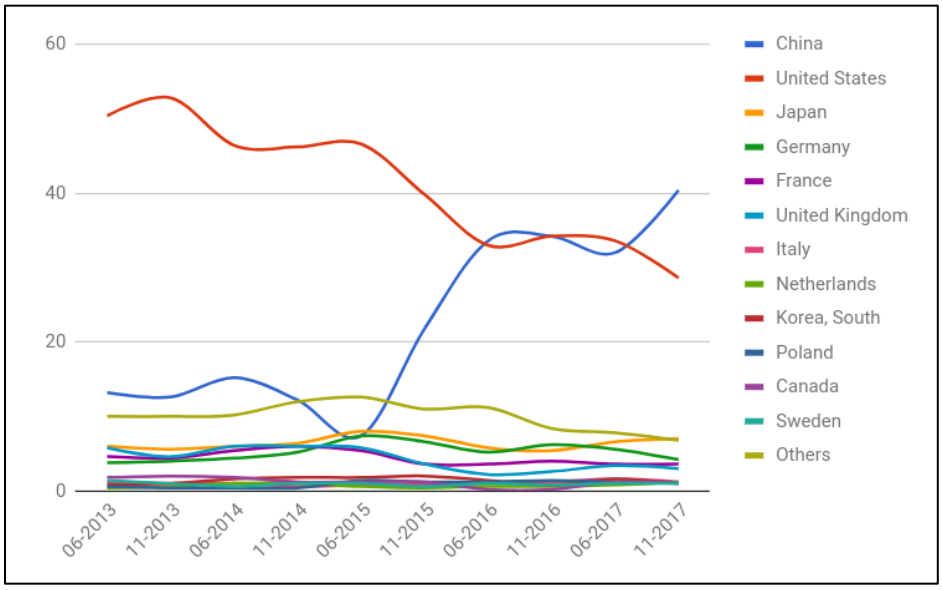

Figure 3. Countries system share over time (TOP500).

When the Rmax values are considered during the last five years as shown in Figure 4, a similar picture seen in Figure 3 is apparent, with China showing an increase while other countries show a decline.

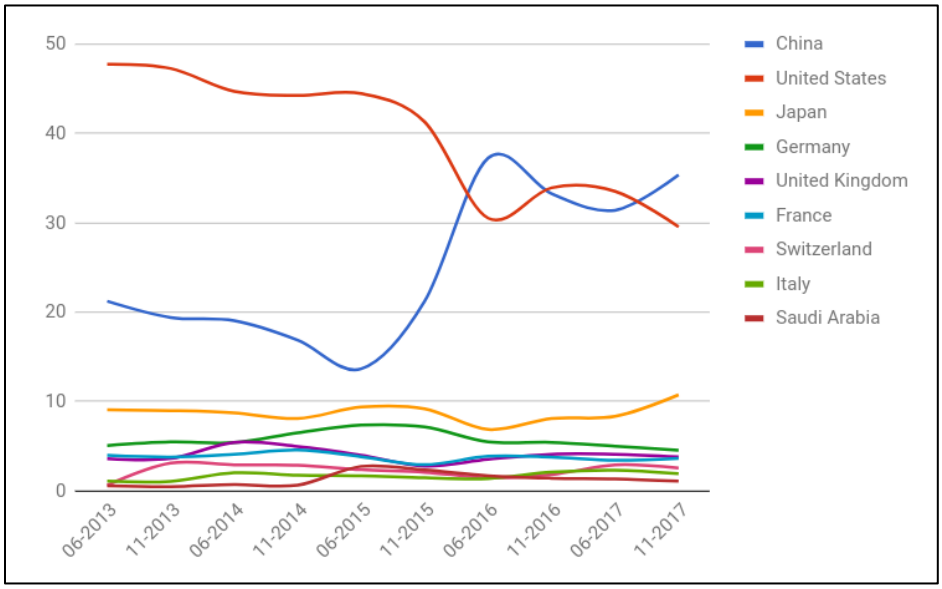

Figure 4. Percentage of cumulative Rmax values (in GFlop/s) for countries (TOP500). Y-axis represents the percentage of Rmax values.

Table 2 shows the number of systems operating in the leading countries. The data included in this table are taken from 2016 and 2017 TOP500 lists. This table also clearly shows that China

overtakes the USA in terms of system share according to the declared TOP500 list in November 2017. Furthermore, the number of systems in European countries is lower.

Top 500 Systems	China	USA	Japan	Germany	France	UK
Nov. 2017	202	143	35	21	18	15
Nov. 2016	171	171	27	31	20	13

Table 2. Leading countries systems shares in the TOP500.

Figure 5 shows the same graph as Figure 4, but excludes the countries outside Europe. The decline in relative performance share is apparent; according to Figure 5, the share of Germany is in decline, but it is still in the lead among the European countries.

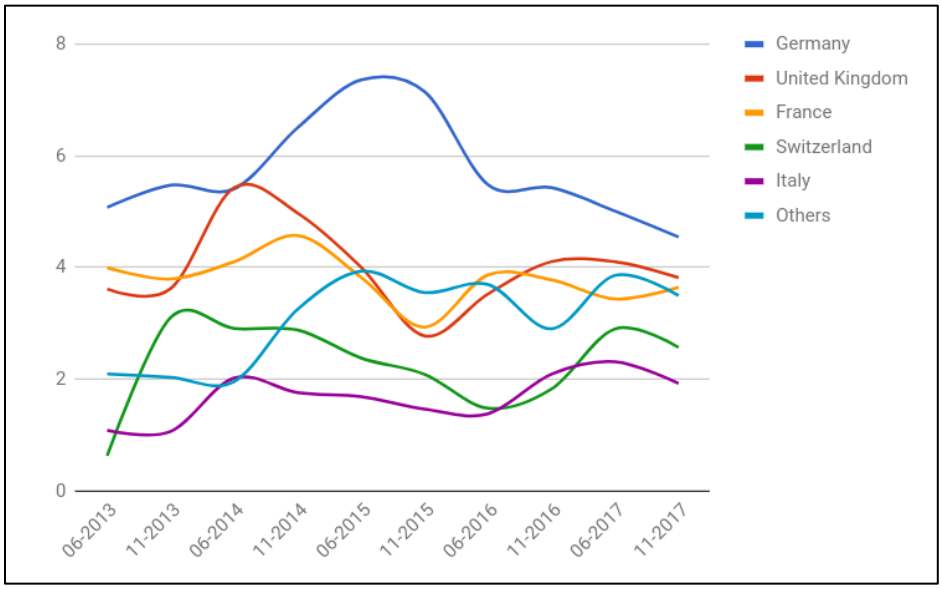

Figure 5. Percentage of cumulative Rmax values (in GFlop/s) for European countries.

Figure 6 and 7 show the presence of European countries in the top 10, 20 and 50 entries of the TOP500 lists released over the last five years. Figure 6 shows the raw values, but the numbers in Figure 7 are normalized according to the size of the sublist.

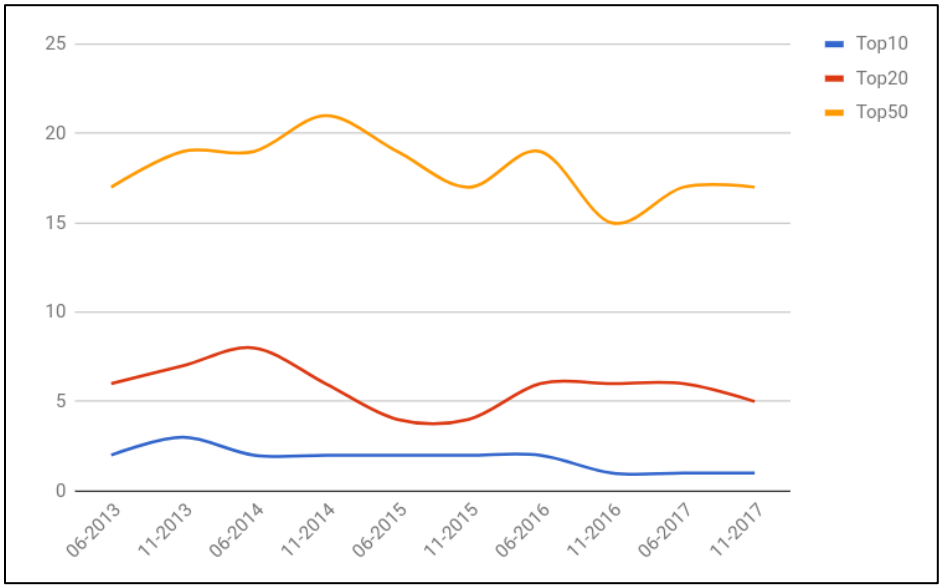

Figure 6. Systems share in Top10/20/50 for European countries

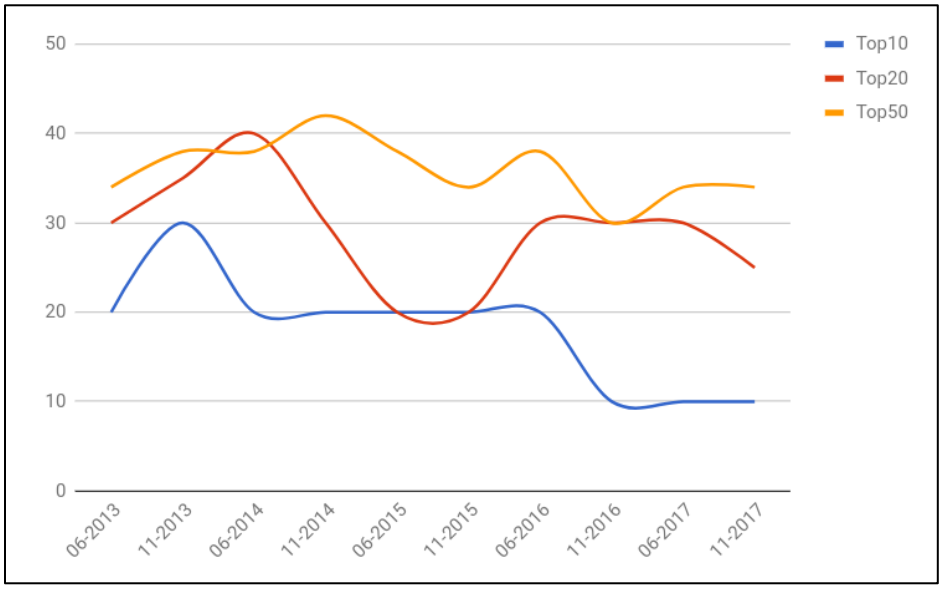

Figure 7. Ratio of systems in Top10/20/50 for the European countries.

The above figures emphasise the decreasing importance of European country HPC systems over the past few years.

2.1.2 Accelerators

Figure 8 shows the fraction of systems equipped with accelerators in the Top50. This figure indicates that half of the Top50 systems includes an accelerator, with the NVIDIA GPU being the preferred choice followed by the Intel Xeon Phi. The remarkable decline in Intel Xeon Phi after 2015 should be noted.

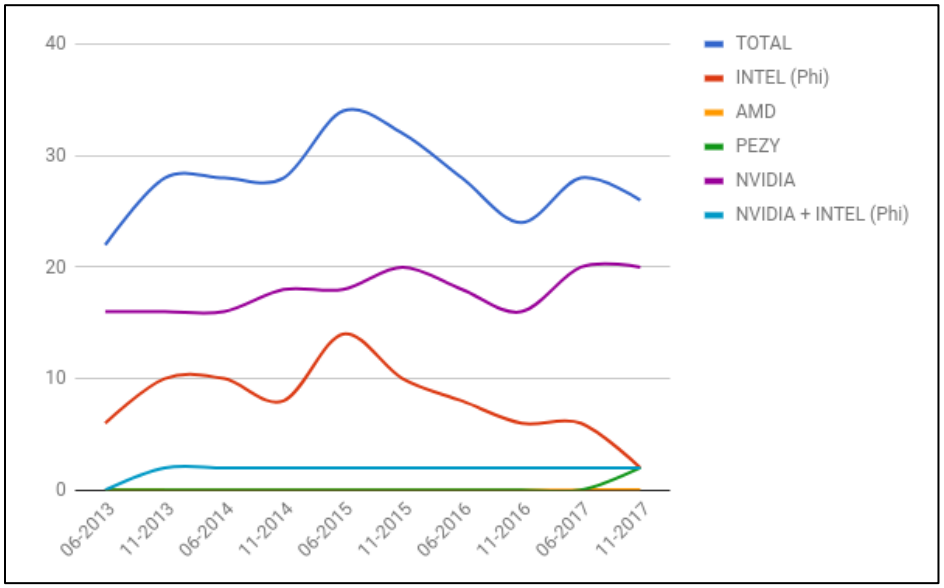

Figure 8. Fraction of systems equipped with accelerators (Top 50).

Figure 9 compares the fraction of accelerators in both Europe and the world based on Top50 and TOP500 rankings. We see that for the November 2017 Top50 and TOP500, the data indicate that only the NVIDIA GPU dominates the European Top50, and in fact this is also true at the global level. Figure 9 also shows that approximately 20% of the systems in the TOP500 are equipped with an accelerator. This percentage is reduced to roughly 3% in the case of the European500.

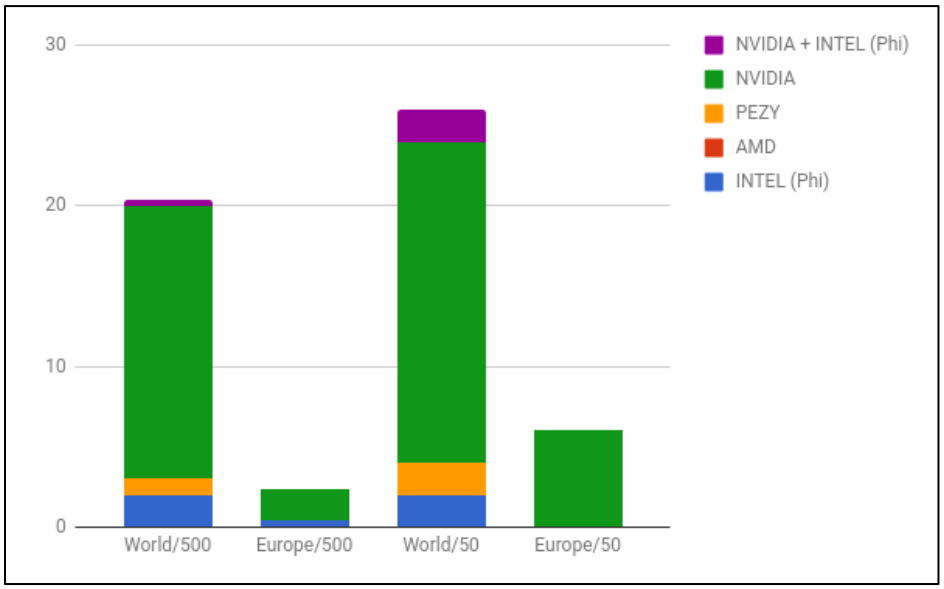

Figure 9. Fraction of systems equipped with accelerators (November 2017).

2.1.3 Age

Figure 10 shows the age of systems (in terms of time of presence in the TOP500) globally and for Europe. The age of systems has been steadily increasing for the last 5 years for the Top50 and TOP500 for both Europe and the world.

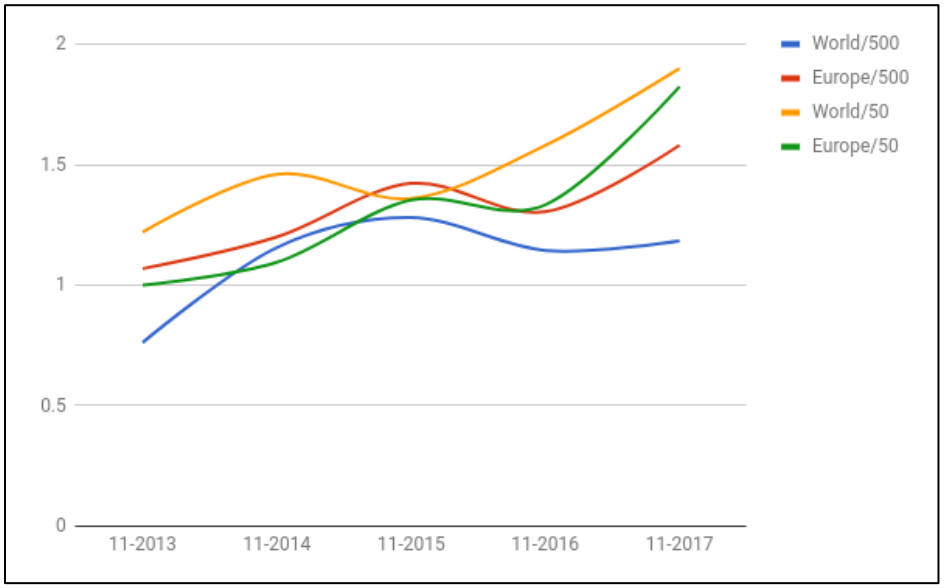

Figure 10. Average age of the systems.

2.1.4 Vendors

Figure 11 and 12 shows the position of vendors globally and in Europe. Globally, CRAY is the dominant vendor with 18 systems. The number of IBM systems, on the other hand, has been steadily decreasing especially starting from 2013, and this decrease has not been affected by the merger with LENOVO. The picture seen in Figure 11 is to some extent also valid for Europe (see Figure 12).

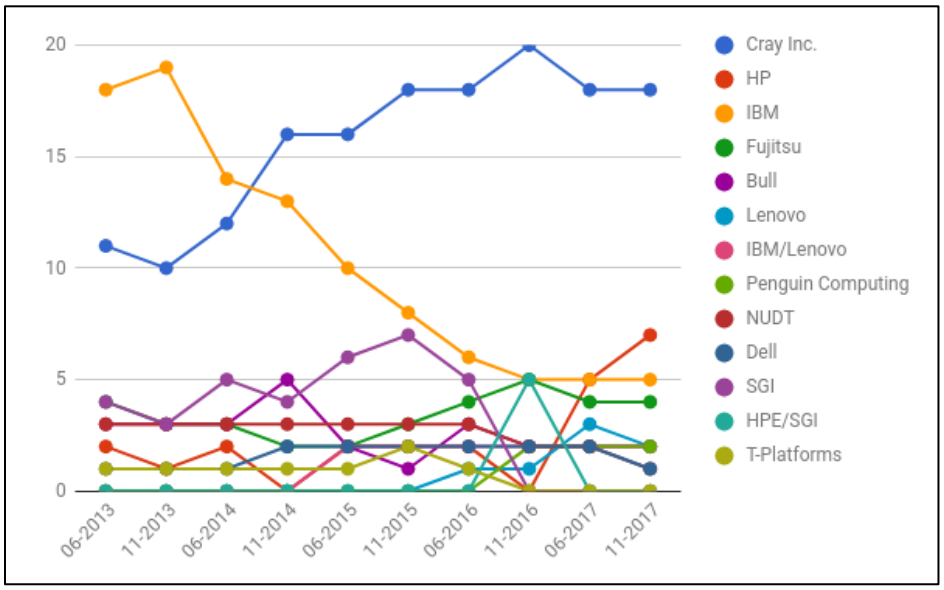

Figure 11. Top50 vendors (world).

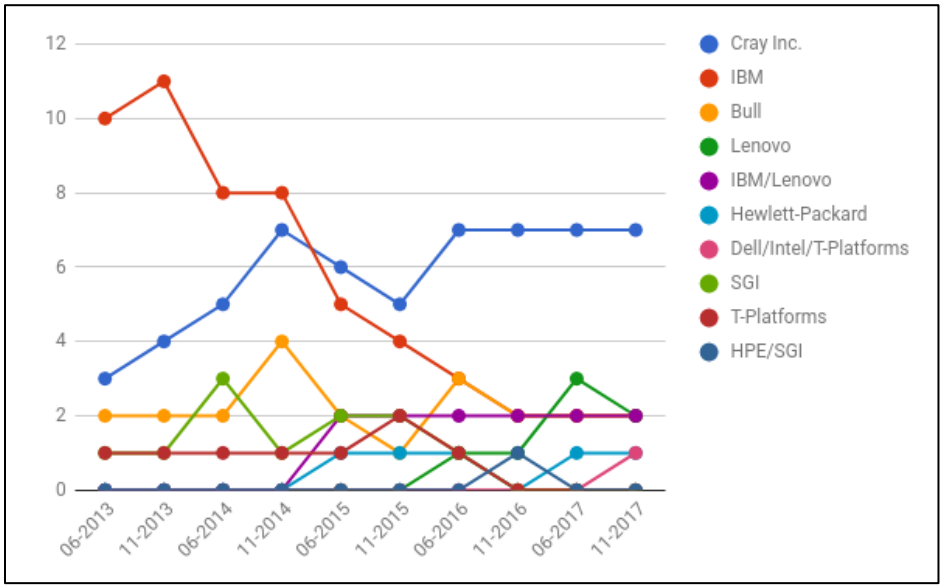

Figure 12. Top50 vendors (Europe).

Figure 13 shows the performance of the European vendor Bull. In the Top50, most of the Bull systems are installed in Europe and the number of Bull systems shows a wavy motion with a recent decrease. For the Top100 (see Figure 14), the number of Bull systems starts to increase in 2014 and it reaches a maximum around 2016 - the number of systems both globally and in Europe is almost identical. When the TOP500 is considered (see Figure 15), the increase in the number of Bull systems around 2016 is clearly visible. Even if after 2016 there is a slight decline Bull's rank in the TOP500 is on the whole increasing (see Figure 16).

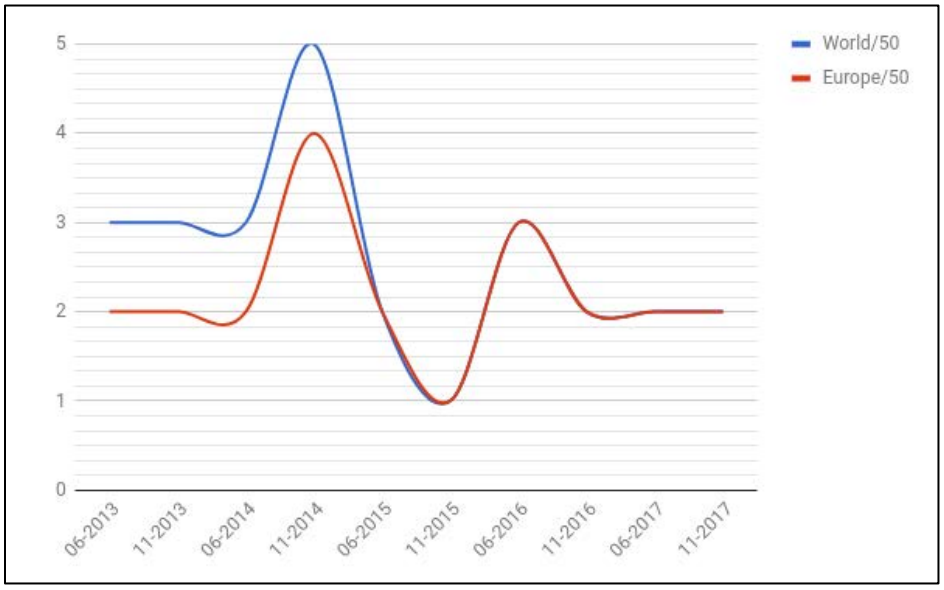

Figure 13. Top50 number of Bull systems.

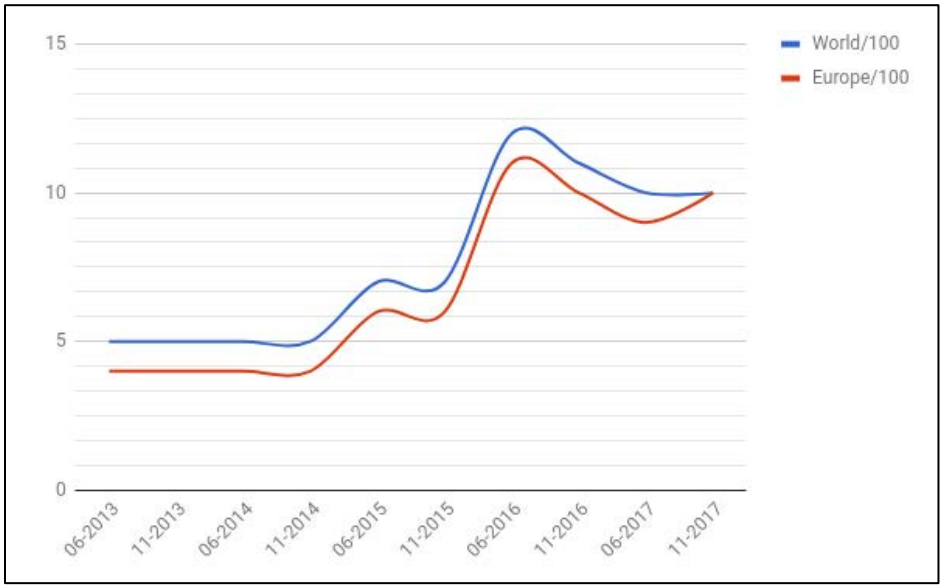

Figure 14. Top100 number of Bull systems.

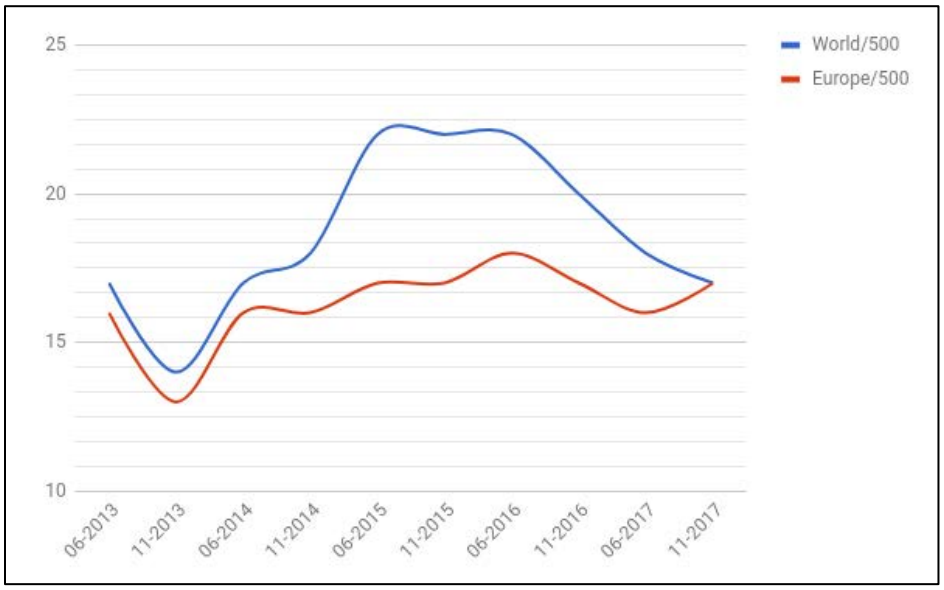

Figure 15. TOP500 number of Bull systems.

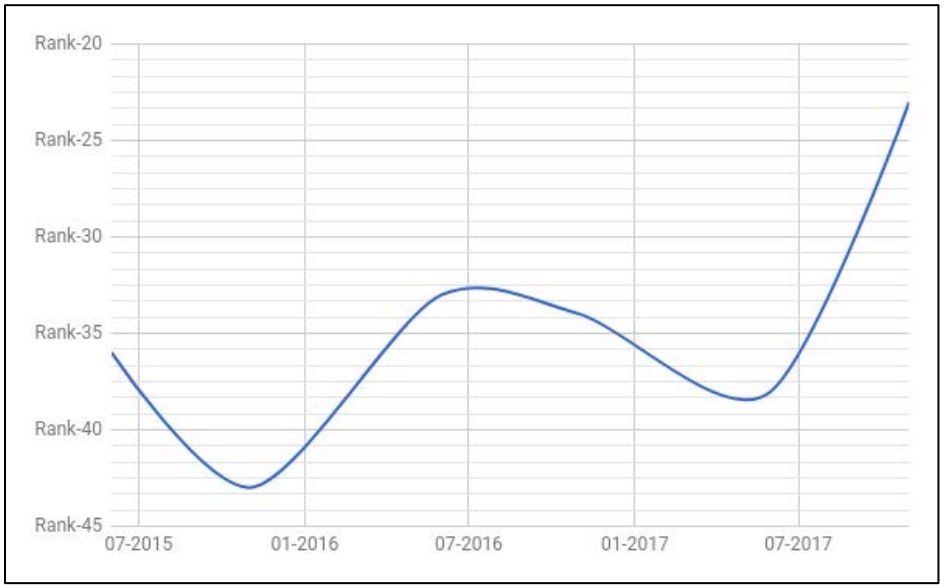

Figure 16. The rank of Bull in TOP500.

2.1.5 Computing efficiency

Figure 17 shows a comparison between HPL and HPCG efficiencies for the first 50 systems. The data show that only 5 systems (namely, K-computer, iDataPlex, Earth Simulator (NEC), Oakleaf FX and the Earth Simulator (IXS NEC), ranking 1st, 35th, 36th, 37th and 38th, respectively) are above 5% efficiency for the HPCG benchmark. The efficiency values for the HPL benchmark however, show a wide degree of variation.

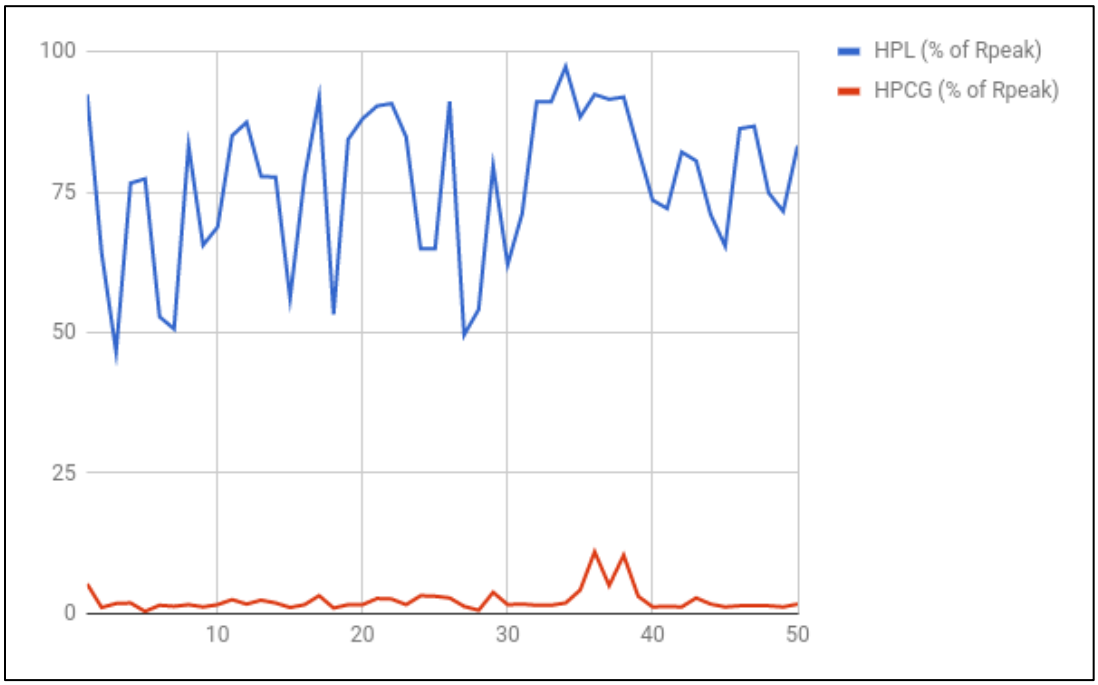

Figure 17. HPL vs. HPCG efficiency comparison.

2.1.6 Energy efficiency

The energy efficiencies of the Top10 and Top50 systems and their reen counterparts are shown in Figure 18 and 19, respectively. In these graphs GFlops/W is used to measure the energy efficiencies. Both figures indicate that not only are the systems in the green list much more energy efficient compared to the Top10 and 50 systems of the TOP500 list but that each year an increase is observed for the energy efficiency. However, this increase is dramatically sharp starting from 2016. In particular, the energy efficiency in 2016 increased from approximately 5 to 11 GFlops/W in 2017.

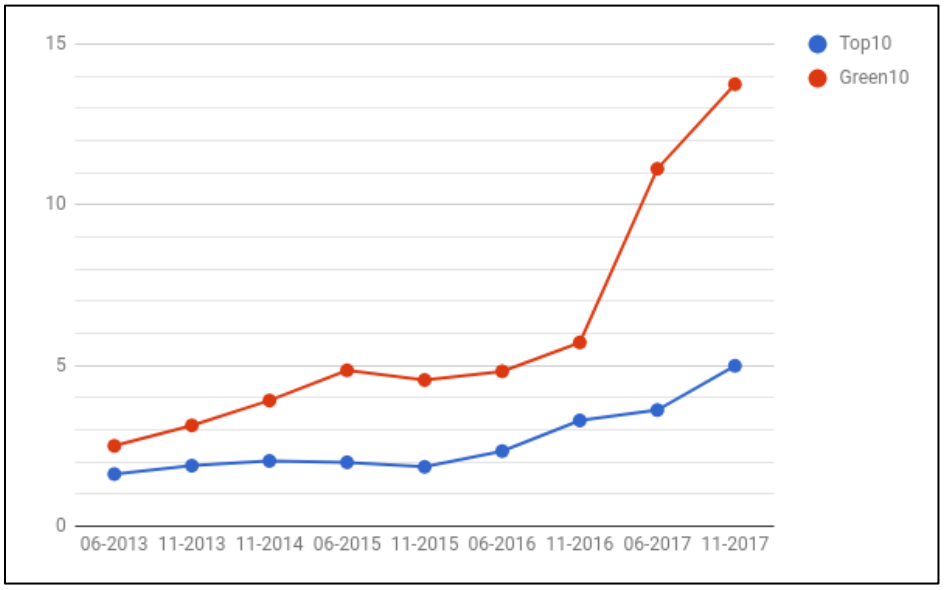

Figure 18. Average energy efficiency in Top10 and Green10.

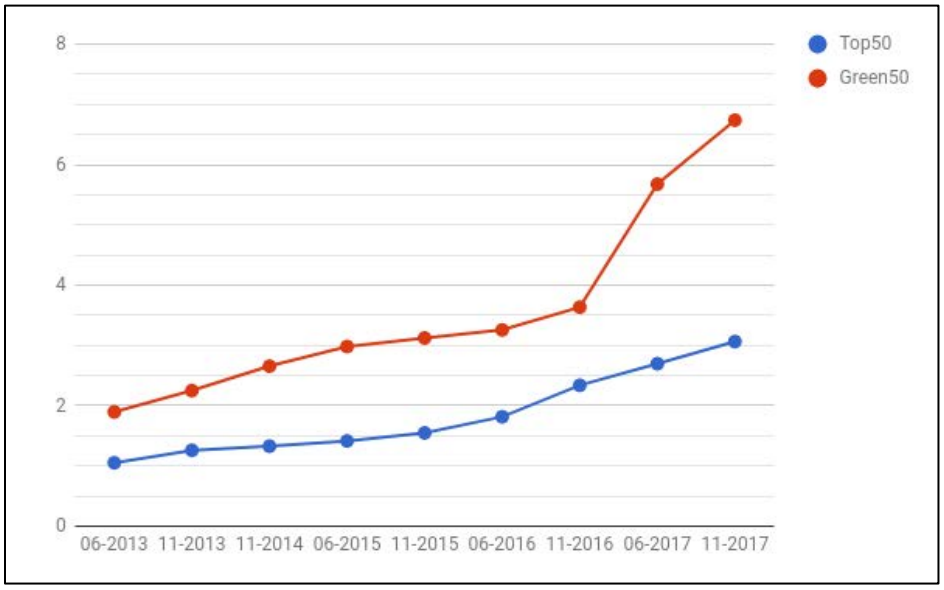

Figure 19. Average energy efficiency in Top50 and Green50.

2.2 Update on Exascale initiatives

In this short overview, we present an update on the high-level plans and the underlying technologies with regards to the exascale strategies in Europe and other regions (see previous PRACE deliverables [4, 5]). Outside Europe, three ecosystems dominate the international HPC scene - China, USA and Japan. Other countries have recognised the importance of HPC and have initiatives in place to develop their ecosystems, but only these regions have global plans encompassing the whole value chain from technologies to usage. There is also a clear trend to create synergies between HPC programmes, Big Data and artificial intelligence initiatives.

2.2.1 Exascale plans in China

It is undeniable that the progression of China in HPC technology and applications during the last years has been impressive and is not slowing down andthe TOP500 ranking gives clear evidence of this trend. In November 2017, the 50th TOP500 list had China overtaking the US in the total number of ranked systems by a margin of 202 to 143. It is the largest number of supercomputers China has ever claimed in the TOP500 ranking, with the US presence shrinking to its lowest level since the list's inception 25 years ago [6]. In the field of HPC technology, the worldwide top HPC system is based on a processor developed in China (the Sunway TaihuLight with a TOP500 number one ranking for the fourth time in November 2017). As a result, China now masters HPC processor and interconnect technology. With respect to applications, even if the Sunway TaihuLight system memory subsystem is not balanced compared to its processing power, Chinese teams won the 2016 and 2017 Gordon Bell awards, achieving impressive simulations with this system.

China has set the clear objective to achieve exascale, and different competing projects are underway (e.g. the Sunway successor, the Tianhe successor and an industrial project led by Sugon) to achieve this.

Tianhe-3 might be the first exaflop/s system, with a prototype expected in 2018 (a few PFlop/s) with the full system expected in 2020; the system will be hosted by the National Computer Centre at Tianjin. However, not much has been unveiled regarding the processor architecture, interconnect technology, core count, node count or the OS, although a likely plan would be to use successors of Sunway's Chinese-designed SW26010 manycore 64-bit RISC processors, together with a custom-designed interconnect.

China is not only aiming to reach exascale as a symbolic target, but more generally developing all the ecosystem [122][123]. There are massive investments planned in several main supercomputing centres, in CNGrid high performance computing network, in middleware, in applications in research, industrial and societal areas, and in education to develop skills for efficient HPC and Big Data applications.

2.2.2 Exascale plans in Japan

In Japan, the most important effort is the post-K project (aka Flagship 2020 [7]) which plans to deliver an exascale class supercomputer – originally planned for 2020, its common use is now rather foreseen around 2021. The project is managed by RIKEN and includes the development of

a new system architecture by Fujitsu, the delivery of a complete software stack and some advanced optimisations in nine application domains.

Fujitsu has developed a new processor-based on the ARMv8 instruction set architecture, collaborating closely with ARM and contributing to the development of the HPC extensions (called SVE) for ARMv8-A, a cutting-edge ISA optimized for a wide range of HPC, meant to increase the performance of the system. Post-K will also rely on a 6D mesh/torus interconnect, which is the successor of K's Tofu.

The system software stack for post-K is being designed and implemented with the leverage of international collaborations: CEA, DOE Labs, and JLESC (joint lab between NCSA, INRIA, ANL, BSC, JSC and RIKEN). The software stack developed at RIKEN is open source and also runs on Intel Xeon and Xeon phi.

Besides the post-K project, Japan is preparing new systems which will be used both for traditional HPC applications and big data applications. Japan has launched an ambitious plan in AI (more than $1B) and the HPC community is working on how to leverage HPC technologies for this field [8].

2.2.3 Exascale plans in the USA

In July 2015, an Executive Order of President Obama established the National Strategic Computing Initiative (NSCI) and gave the US HPC a high political visibility [9]. Due to China's development, the US HPC leadership has been jeopardised if not reduced, and there are expectations and attempts to keep or regain the lead (clearly the US lost already the 'TOP500' symbolic leadership in terms of number of systems, even if they are still leading in terms of technology and market presence).

It is expected that they will regain the lead in HPC systems with the installation of three top systems issued from the Coral procurement (Coral = Collaboration of Oak Ridge, Argonne, and Livermore). Two systems (Summit in Oak Ridge National Lab [10] and Sierra in Lawrence Livermore National Lab [11], both based on the OpenPOWER architecture + NVIDIA GPUs) were installed in 2017 and are planned for production in 2018 and will deliver between 125 and 200 PFlops each. The third system (Aurora), based on the Intel KNH architecture and originally planned to be installed in 2018 in the Argonne National Laboratory, will now instead be a re-designed exascale system which will be delivered in 2021 [12].

The main efforts for the delivery of the exascale computing capability are now organised under the Exascale Computing Project (ECP [13]) managed by the Department of Energy (DoE). The ECP is a collaborative effort of two U.S. Department of Energy organizations: the Office of Science (DOE-SC) and the National Nuclear Security Administration (NNSA). The ECP covers the development of technologies for both hardware and software, systems and applications and procurements of exascale systems will follow SC and NNSA processes and timelines. ECP emphasises the concept of 'capable exascale computing ecosystem', focusing on real and sustainable applications acceleration – both from the computational and data analysis capabilities – and not on mere simplified benchmarks like LINPACK.

ECP has four focus areas:

- Application development;

- Software technology;
- Hardware technology;
- Exascale systems testbeds.

The first set of research contracts has been awarded by ECP for hardware technologies (six projects), software technologies (thirty-five projects), co-design centres (fifty-five projects) and applications (twenty-three projects).

2.2.4 Exascale plans in Europe

HPC is considered as one of the key contributors to the Digital Single Market (DSM) strategy announced by the EC in April 2016, which confirms and widens the scope of the 2012 EC strategy [14]. The European Cloud initiative [15] aims to provide researchers, industry, SMEs and public authorities with access to world-class supercomputers, thus unleashing their innovation and transformation potential. This includes the notions of a European Science Cloud for scientists to access an underlying rich infrastructure with computing, data storage, access and processing as well as networking capabilities – the European Data Infrastructure (EDI). HPC is thus considered essential for the European Data Infrastructure and for the European Open Science Cloud as it will provide the capacity to analyse vast amounts of data, in addition to providing high end computing capabilities.

The reader should also be reminded that the contractual Public-Private Partnership on HPC (cPPP on HPC) had predated this initiative: it entered into force in January 2014 to develop an ambitious RI strategy [16], supporting HPC applications and technologies development, in addition to PRACE which had started as a pan-European supercomputing infrastructure in 2010. The current status of the HPC cPPP projects is described in Section 2.4.

The European High-Performance Computing Joint Undertaking (EuroHPC JU) is a new entity planned to start in 2019 to pool European resources to fund a pre-exascale supercomputing infrastructure in 2019-2020, then to develop exascale supercomputers based on competitive EU technology that the Joint Undertaking could acquire around 2022/2023. The EuroHPC Joint Undertaking builds on the declaration launched in Rome in March 2017. Fourteen Member States have now joined EuroHPC, after seven countries initially signed this declaration. EuroHPC JU will enable Member States to coordinate together with the Union their supercomputing strategies and investments, and will likely progressively take over from the HPC cPPP for the RDI coordination.

The EuroHPC budget of EUR 486 million will come from the present budgetary framework of the Union, and used for both Horizon 2020 and Connecting Europe Facility (CEF) programmes. The budget is expected be matched by a similar amount from the participating countries and private entities should also provide in-kind contributions. The Joint Undertaking will provide financial support in the form of procurement or research and innovation grants to participants following open and competitive calls. The HPC EU investment up to the end of 2020 is thus close to EUR 1 billion while another EUR 4 billion are foreseen under the next financial framework.

The Commission is also launching an ambitious flagship initiative to unlock the full potential of quantum technologies, including quantum computing and communications.

National initiatives include (but are not limited to) such plans as:

- Germany's BMBF "Smart Scale" accelerated HPC investment [120], involving Gauss Centre for Supercomputing three centres (FZJ HLRS and LRZ), with new supercomputers at each of them, and furthered education and training programs
- France's Plan Supercalculateurs launched in 2014 [121] with updated targets and renewed support in 2016, with an exascale technologies facet but also support for software applications, and actions for the development of HPC use in industry and SMEs.

2.3 Business analysis

HPC Business analyst Timothy Prickett Morgan [17] predicts that it will take two decades for HPC to morph into AI. However, HPC as we know it is still an important driver of innovation, vital to the global economy. Evidence for this has been provided by Intersec360 Research and Hyperion Research when they revealed their statistics and predictions at the ISC 2017 Conference. According to Intersec206 [17], 2016 represented the seventh consecutive year that the HPC market grew, but the growth has slowed. In 2016, the market grew by 3.5% to reach $33.59 billion in sales across all products and services relating to HPC.

Cloud HPC revenues grew at 6.4%, but revenues hit $784 million across all cloud infrastructure makers in 2016 (see Table 3). HPC software accounted for $8.91 billion in revenues and rose by 3.5%, while services comprised another $3.82 billion, up by 1.4% for 2016 [17].

	2015	2016	Change	Growth
Servers	11,103	11,471	369	3.3%
Storage	5,503	5,778	274	5.0%
Services	3,770	3,824	54	1.4%
Software	8,606	8,910	304	3.5%
Networks	2,678	2,767	89	3.3%
Cloud	737	784	47	6.4%
Other	1,987	2,053	66	3.3%
Total	34,384	35,587	1,203	3.5%

Table 3. Total HPC Revenue by Product Class (in million $)

Looking further ahead to 2021, Intersect360 believes that the market will grow a little faster in the years between 2016 - 2021, with a compound annual growth rate of 4.3 percent and attaining a level of $43.94 billion in spending. At the same time, Intersect360 is not particularly optimistic about HPC in the cloud.

	2016	2021	CAGR 2016-2012
Server	11,200	14,819	5.8%
Storage	4,316	6,269	7.8%
Middleware	1,277	1,786	6.9%
Applications	3,739	5,071	6.3%
Service	1,907	2,309	3.9%
Total Revenue	**22,439**	**30,253**	**6.2%**

Table 4. Hyperion Research Market Forecast on the Broader HPC Market ($ Millions).

The HPC Group of Hyperion Research is just as optimist as Intersect360 about HPC sales looking to the future, and expects the market to expand by a third between 2016 and 2021 [17]. What is clear from Table 4, is that this growth is not just due to pre-exascale and exascale systems at the top HPC centres. The growth that Hyperion is projecting is projected across all classes of machines. It is also pretty optimistic about other aspects of the HPC market, including storage, middleware, applications, services and servers. Intersect360 believes that the HPC market is considerably larger than what Hyperion thinks it is, and that comes down, we believe, to the fact that people have different opinions about what constitutes HPC and what does not.

For 2016, the revenues for the broader European HPC Market are shown in Table 5.

	2016
Server	3,551,920
Storage	1,279,344
Middleware	375,141
Applications	1,092,638
Service	570,586
Total Revenue	**6,869,630**

Table 5. Revenues for the Broader European HPC Market ($ Thousand)

2.4 EU HPC Projects Landscape and PCPs

European HPC in Horizon 2020 is based on three pillars (Figure 20) [18, 19]:

- Research Infrastructure (including PRACE - Partnership for Advanced Computing in Europe);
- HPC Technology (represented by ETP4HPC [20]) ;
- Application Expertise (developed through the Centres of Excellence in Computing Applications - CoEs [21]).

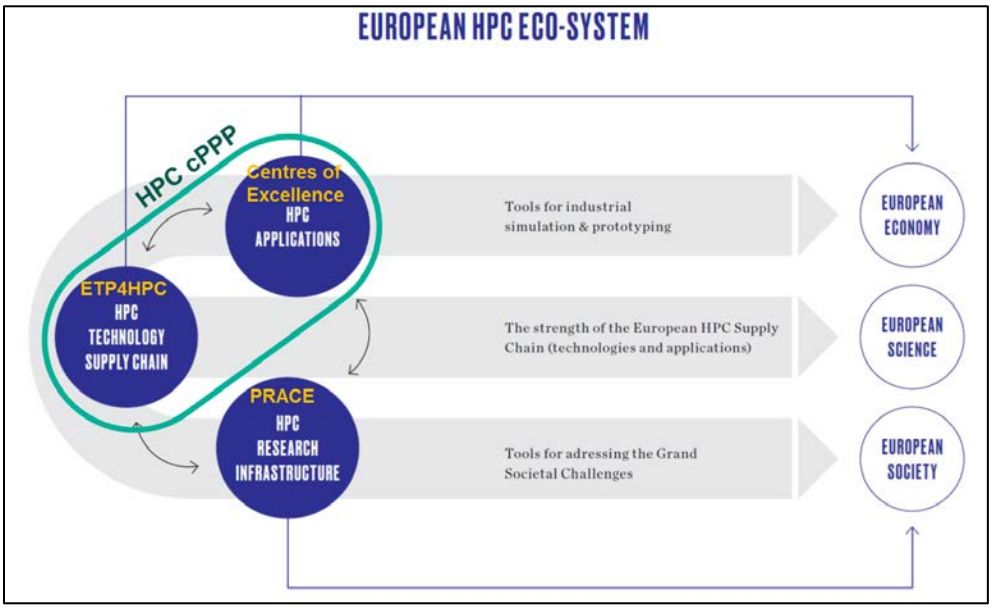

Figure 20. European HPC in Horizon 2020

In this section, we briefly highlight the Technology and Applications pillars of the ecosystem, which form the contractual Public-Private Partnership (cPPP) with the European Commission [22].

The FET-HPC part of the programme concerns the development of basic HPC technology, while the CoE sub-programme supports Centres of Excellence in Computing Applications, consolidating the European HPC application expertise. The ecosystem development on the other hand is supported by a series of Coordination and Support Actions, which orchestrate the European HPC strategy. It is important to note that some relevant elements of the European HPC effort might fall into other (non-HPC) programme parts such as LEIT [23]; centres of Excellence fall into EINFRA part of H2020 Pillar 1.

The cPPP-related calls for projects are mostly fed and influenced by the HPC multi-annual roadmap developed by ETP4HPC (Strategic Research Agenda [24], periodically updated to serve as reference for the successive Work Programmes: 2014-2015, 2016-2017, and now 2018-2020 starting).

Until now, four calls have been closed under Work Programmes 2014-2015 then 2016-2017 (Table 6). The total amount of committed H2020 funding is 219.5 M€. This should be compared to the cPPP total provisioned budget of 700 M€: this means the largest fraction of funding is still to be granted under Work Programme 2018-2020 calls.

	WP	Topic	Type of Action	Call Deadline	Start
Documented in Handbook 2017	FETHPC 2014	HPC Core Technologies, Programming Environments and Algorithms for Extreme Parallelism and Extreme Data Applications HPC Ecosystem Development	19 RIA 2 CSA	Nov. 2014	Sept. 2015
	FETHPC 2016	Co-design of HPC systems and applications	2 RIA	Sept. 2016	Q2 2017
	CoEs 2014 (EINFRA)	Centres of Excellence for Computing Applications	9 RIA	Jan. 2015	Sept. 2015
	FETHPC 2017	Transition to Exascale Computing HPC Ecosystem development	11 RIA 2 CSA	Sept. 2017	Q2 2018

Table 6. The FETHPC and COE calls in H2020

The European HPC Handbook [25] includes up-to-date details of the European HPC Technology and Application Projects within the European HPC ecosystem; 2015 and 2016 Handbooks can also be consulted. The 2017 handbook covers the first 3 calls mentioned in Table 6 -the most recent call of September 2017 led to the selection of 11 more RIA and 2 more CSA projects, which are about to start in Q2 of 2018.

In Figure 21, the RD technology projects chart layout indicates their main research area, relating to the ETP4HPC SRA topics.

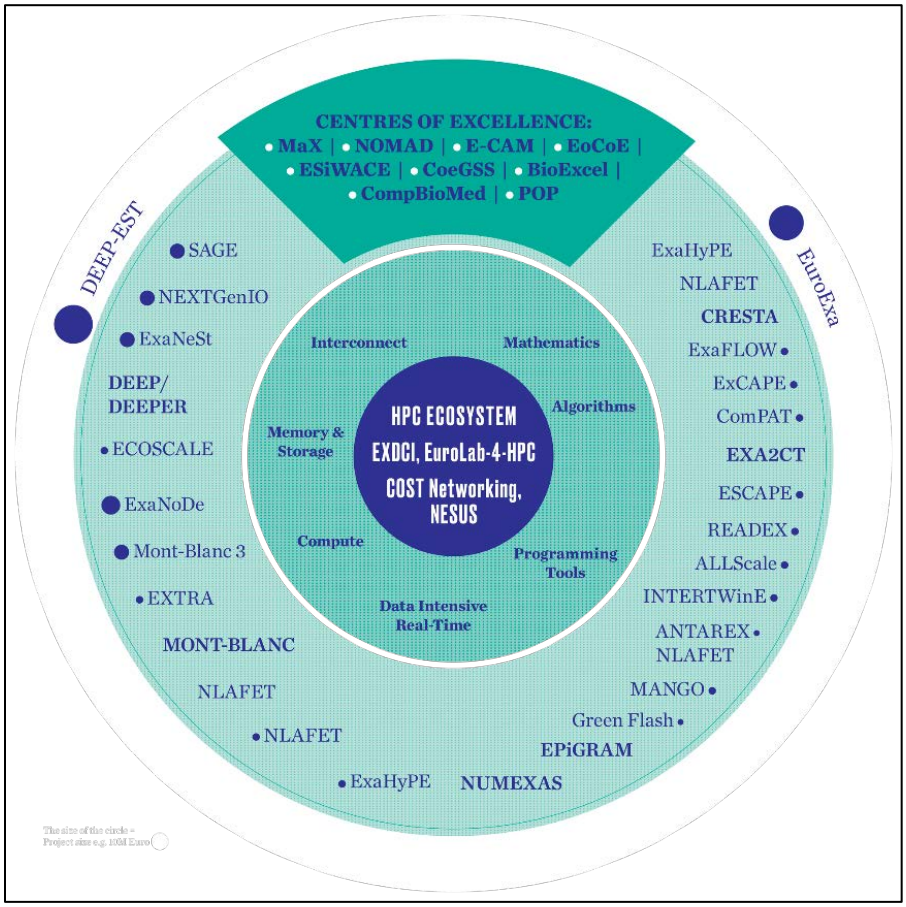

Figure 21. Portfolio of H2020 HPC projects – Technology and applications R&D

FP7 has also funded some technology R&D efforts r via PRACE-3IP and HBP PCPs (Pre-Commercial Procurements).

PRACE-3IP PCP

The PRACE-3IP PCP opened a call for tender in November 2013 with Phase III ending in December 2017. For this final phase of the PRACE-3IP PCP, which began in October 2016, the following suppliers were awarded a contract:

- ATOS/Bull SAS (France), “Frioul” hosted at CINES (France);
- E4 Computer Engineering (Italy) “D.A.V.I.D.E.” hosted at CINECA (Italy);
- Maxeler Technologies (UK), “JUMAX”, hosted at JSC (Germany).

During this final phase the target compute capability was around one PFlops. Initially, two were supposed to be awarded in phase III, but the GoP committee chose to select three, for the same global budget thus leading to solution resizing.

The contractors have deployed the pilot systems to demonstrate technology readiness of the proposed solution and the progress in terms of energy efficiency, using high frequency monitoring designed for this purpose, and bring the proof of extensibility to 100 PFlops. The contract for each supplier comprised of performance application commitments (4 from UEABS and HPL) in terms of Time-to-Solution and Energy-to-Solution. Access to these systems was granted to PRACE

partners, after the PCPs team performed their evaluations. One of the PRACE-4IP WP7 extension topics was the EUABS energy profiling and optimisation.

Due to very late deployment, the JUMAX system could not be assessed by PRACE-4IP WP7, and more work was carried out on Frioul and D.A.V.I.D.E. During the final review, the EC asked for a follow up (on going within PRACE-5IP WP7) for a more complete evaluation and a global restitution workshop of PCP outcomes that could take place early in 2018.

HBP PCP

The HBP PCP finished on 31st January 2017. To ensure the availability of the HBP High-Performance Analytics and Compute Platform (HPAC), the project published a tender for a PCP in April 2014 focussing on R&D services in the following areas: integration of dense memory technologies, scalable visualization and dynamic management of resources required for interactive access to the systems. In phase III, Cray and a consortium consisting of IBM and NVIDIA were selected. These contractors implemented their proposed solutions and evaluation is ongoing since the pilot systems were installed in summer 2016. JUELICH will continue to keep the pilot systems in operation to keep the solutions available to the HBP project via the HPAC.

PPI4HPC

The Public Procurement of Innovations for High Performance Computing (PPI4HPC) project [26] is funded under H2020 Call EINFRA-21-2017: Platform-driven e-infrastructure innovation, topic (a) Support to Public Procurement of innovative HPC systems, PPI [27].

In PPI4PHC, a group of leading European supercomputing centres formed a buyers group to execute a joint Public Procurement of Innovative Solutions (PPI) for the first time in the area of High-Performance Computing (HPC). The co-funding by the European Commission (EC) will allow for a significant enhancement of the planned pre-exascale HPC infrastructure from 2019 and pave the path for future joint investments in Europe – such as under the EuroHPC umbrella. The total investment is planned to be about € 73 million, including a €26 million EC contribution.

The partners involved, namely BSC, CEA/GENCI, CINECA and JUELICH, work together on coordinated roadmaps for providing HPC resources optimised to the needs of European scientists and engineers. The decision on which innovative solutions will be procured at the different sites will be made following these roadmaps, but the final decision will remain with the individual sites.

The first concrete steps were:

- PPI4HPC organized an Open Dialogue Event on 6 September 2017 to inform the market of the future joint procurement and to gather input from the market.
- A next step was to hold one-to-one meetings with vendors in order to investigate HPC solutions for the supercomputing centres in the PPI4HPC project. The project partners held 15 one-to-one meetings with major HPC companies including various SMEs in September and October 2017. The aim was to have in-depth technical discussions, while vendors also had the opportunity to outline future HPC solutions, potentially able to fulfil the needs of the PPI4HPC partners.

2.5 Cloud computing and virtualization

2.5.1 Overview of current trends in HPC clouds

HPC targeted cloud services have already been available for some time. Players like Amazon (AWS), Microsoft (Azure), as well as T-Systems (HPC Cloud) offer services that target the HPC market. OpenStack is also gaining popularity within research organizations, and several traditional HPC centres already offer a solution on top of OpenStack. These generally complement, rather than replace, the traditional HPC systems.

2.5.2 Commercial cloud vendors

The support for HPC resources in the commercial cloud vendor space has had strong growth in 2017. Several vendors have their targeted HPC offering with various levels of service built on top of it.

2.5.2.1 T-Systems HPC cloud

In 2017, T-Systems released their Huawei built HPC cloud offering under Open Telekom Cloud. This includes HPC sized virtual machine flavours with Infiniband, bare metal servers, GPU servers and access to shared file storage.

A subset of the high performance flavours, the 12-core variants, include Infiniband EDR support. The HPC flavours scale to 32 cores and 256GB RAM while the GPU flavours are based on the NVIDIA M60 product, which is targeted more towards the workspace market compared to other offerings with NVIDIA P100 GPU cards.

As a summary, the T-Systems services look like a good start, but it will most likely not satisfy the more demanding HPC users. The lack of Infiniband support in larger flavour sizes, and the lack of computationally performing GPUs rule out some heavier workloads.

2.5.2.2 Amazon AWS

On the hardware side, there were no big surprises in Amazon's offerings in 2017. As expected for the largest cloud player, they were among the first ones to release new hardware when it came out. Notable mentions are new Intel “Skylake” based virtual machines, now with 25 Gbit Ethernet connectivity. While the Ethernet connectivity is faster, it is unlikely to directly compete with Infiniband-enabled cloud services. Amazon was also early in introducing the new NVIDIA Tesla V100 accelerator cards.

Amazon released a new hypervisor called "Nitro" in 2017 where much of the workload which traditionally was done in software is shifted to custom ASICs. This should reduce the overhead in fields where virtualization has traditionally suffered the most, i.e. IO and networking. In addition to this, Amazon also released bare metal servers as a service in 2017.

On the management side, Amazon introduced NICE EnginFrame, an HPC orchestration service designed for cluster and cluster workload management.

2.5.2.3 Azure

Many HPC specific developments were announced in 2017 for Microsoft's Azure cloud. Azure Batch was updated with support for AI workload management and Singularity containers, the latter allowing use of shared computing environments while maintaining access to specialized HPC hardware such as high-speed interconnects or GPUs. Hardware-wise, the new NVIDIA Tesla V100 accelerators were added in selected regions.

Use of RDMA is consistently more visible, in sync with port speeds increasing from 10Gbps to 25Gbps and above. In the cloud economy, CPUs should be prioritized to serve customer workloads rather than high throughput TCP traffic and this is enabled by RDMA. RDMA over Infiniband is Azure's documented go-to solution - for example for MPI jobs. RDMA over Converged Ethernet (RoCE) remains integral to Azure, but its focus is still more in storage and other cloud backend traffic rather than HPC jobs. Resiliency for packet loss has only recently started to find its way to mature RoCE implementations. On the scale of Azure, it may be feasible to carry out bespoke flow control/congestion control to support RoCE deployments. It remains to be seen if the same applies for smaller HPC cloud providers, who typically run either lossless Infiniband or lossy Ethernet networks.

In the summer of 2017, Microsoft purchased Cycle Computing. A central aspect of Cycle's offering is to ease the management of hybrid HPC deployments. Microsoft also announced a partnership with Cray. This effectively means that they will operate dedicated Cray supercomputers in Azure datacentres on behalf of given customers, and from there enable a clear integration path towards the cloud computing and data processing facilities available on the rest of the Azure platform. User workloads that involve processing of data on virtualized cloud infrastructure in tandem with bare metal supercomputing are the clear trends here.

2.5.2.4 Google Compute Engine

From the big cloud players, Google Compute Engine (GCE) seems to have a smaller focus on HPC compared to the others. While GCE introduced new hardware quite early in 2017 - specifically Intel “Skylake” CPUs, and NVIDIA Tesla P100 accelerators, the rest of the HPC stack is still lacking. There is little information about interconnect options, so scaling beyond single machines in parallel computation will be a problem with GCE.

Google did partner with Altair to bring their suite of HPC applications to GCE, but in general the focus of GCE seems to be elsewhere.

2.5.3 Open Cloud HPC Front

2.5.3.1 OpenStack

OpenStack has been the largest and most deployed in-house cloud platform for a while. In general, OpenStack has the required functionality to build a HPC cloud, and it has been used in several cases. Using technologies like PCI pass-through, SR-IOV and bare-metal provisioning one can build a HPC service with the desired balance of flexibility and performance. There are few if any

technical limitations for building your own HPC cloud any more. The problem has moved towards a business question of cost/benefit analysis of a HPC cloud service vs. a traditional HPC service. This will widely vary depending on the use case, and there are no clear correct answers for this. As a rule of thumb, the cost of a HPC cloud service is more than a traditional HPC service, but it adds flexibility.

2.5.3.2 Kubernetes

Traditionally, HPC cloud discussions have revolved around IaaS. Lately, this has been changing. The container orchestration engine Kubernetes has developed quickly, and offers cloud services on a higher abstraction layer. Its use of containers allows for an easier mitigation of some cloud performance impacts compared to virtualized clouds. Kubernetes can be used with HPC, but the adoption will most likely depend on how many workloads get ported to the Kubernetes workload paradigm, and how many use traditional HPC batch scheduling methods. For workloads using traditional batch scheduling, tools like Singularity provide the benefits of the containerization.

2.5.4 Use cases

Using commercial cloud providers is a real option for some HPC workloads. In small to medium-sized use cases, where you do not have a long-term need for resources, or the need varies heavily, commercial cloud providers start to have solutions. These infrastructure will still struggle to support high-end use cases, which are normally run on Top 500 machines. In addition, for longer-term use, cost savings will likely be achieved by using a dedicated system.

When deploying new HPC systems, it is worth to at least consider if they should be deployed as an internal HPC cloud (even bare metal), or as a traditional system. Both have their benefits, and the ultimate choice boils down to cost, flexibility requirements, existing in-house knowledge and what workloads must be supported.

2.6 Consolidation in the HPC market

The last two years have brought several acquisitions and mergers in the market that can improve the quality of products for HPC and the vendor market is reshaping due to current IT trends. In the following sections, the main consolidations in the server/storage and semiconductor areas are outlined:

2.6.1 Server and storage

- Hewlett Packard Enterprise acquisition of SGI – SGI's technology including in-memory high-performance data analytics and leading high-performance computing solutions extends the HPE portfolio.
- Hewlett Packard Enterprise acquisition of Nimble Storage – Nimble all-flash and hybrid-flash storage solutions extends the HPE portfolio.
- Cray acquisition of Seagate's ClusterStor – Cray takes over development, support, manufacturing and sales of the ClusterStor product line.

2.6.2 Semiconductor

- Cavium acquisition of QLogic – Qlogic high performance networking infrastructure solutions extends the Cavium portfolio.
- Softbank acquisition of ARM – SoftBank is expected to use the ARM deal to bolster its Internet of Things plans.
- Marvell acquisition of Cavium – Cavium's portfolio of multi-core processing, networking communications, storage connectivity and security solution extends Marvell's storage, networking solutions and high-performance wireless connectivity products.
- Extreme Networks acquisition of Avaya Inc. – Avaya Networking will provide Extreme Networks with a broader set of networking technologies.
- Extreme Networks acquisition of Brocade Communications Systems – technology assets from Brocade including the SLX, VDX, MLX, CES, CER, Workflow Composer, Automation Suites, and certain other data centre related products extends the Extreme portfolio.

3 Core technologies and components

3.1 Processors

This section discusses the recent viable processor technologies and upcoming trends.

3.1.1 x86_64 processors (INTEL/AMD)

3.1.1.1 Intel

Intel is the leader in the server processor market. The company manufactures two processor families targeting the server market, the Xeon and Xeon Phi processors.

The Xeon line is represented by its recent incarnation, code-named Skylake, featuring up to 28 cores (56 threads, when HT is enabled) per socket and it is capable of running up to 3.80 GHz. The Xeon Phi line is represented by the processor code-named Knights Landing and features up to 72 cores in a socket (288 threads, when HT is enabled) running up to 1.7 GHz, for the high-end model 7290. The most popular model (i.e. due to far better performance/price) is the 7250 model (68 cores@1.4 GHz) as it includes 16 GBytes of MC-DRAM memory that provides 400 Gb/s of bandwidth.

Intel manufactures additional products such as high-speed fabric interconnect and FPGAs, but we can also highlight their recent addition into the non-volatile memory drives. Their Intel Optane SSD DC P4800X is the industry-leading combination with high-throughput, low latency, high QoS and high endurance. It combines the attributes of memory and storage aiming to break through storage bottlenecks. It accelerates applications for fast caching and storage, increasing scale per server. Data centres based on the latest Xeon processors can now also deploy bigger and more affordable datasets.

On the 13th of November 2017, Intel announced that they will withdraw any future Intel Xeon Phi processors (code name Knights Hill). They will instead target a new platform and new microarchitecture specifically designed for exascale. This could become the building block of a future platform for Argonne Laboratory Computing Facility, in the context of the CORAL US scientific computing roadmap, moving target from 180 PFlops in 2018 to 1 EFlop in 2021. For short-term HPC projects with current Xeon processors, this would imply the use of Skylake or the successor Cascade Lake Scalable Processor, which will be described in the following sections. For the mid-term perspective, the “manycore low frequency” seems to have been abandoned, but the MCDRAM could be present.

3.1.1.1.1 Skylake Scalable Processors

As of 2017, the latest available micro architecture for server class processors from Intel is the Skylake Scalable Processor series, introduced in July 2017. Skylake represents the architecture step in Intel’s recently adopted Product-Architecture-Optimization (PAO) cycle. It should be noted, that while this is not a process increment step, the server configuration models are fabricated using the

enhanced 14+ nm process, as opposed to the 14 nm process used in their client configuration counterparts.

Key changes from the previous Broadwell architecture, apart from the 14+ nm fabrication process, include the introduction of the Omni-Path architecture, the replacement of Cluster-on-Die implementation with the Sub-NUMA clustering and increased Bus/Chipset bandwidth (8.0 GT/s from 5.0 GT/s in previous models). Also, in addition to various improvements such as larger pipelines, buffers and micro instruction cache, the Level 2 cache is increased to 1 MB per core (up from 256 KB per core). However, the Level 3 cache is reduced from 2.5 MB per core to 1.375 MB per core, and becomes non-inclusive in regard to the other levels of cache. Finally, the Translation Lookaside Buffer (TLB) has also received some improvements, making the Instruction Translation Lookaside Buffer (ITLB) 8-way associative (previously 4-way associative) and Second Level Translation Lookaside Buffer (STLB) 12-way associative (previously 6-way associative). In addition, various improvements increased the throughput of core instructions, e.g. most Arithmetic Logical Unit (ALU) operations and fused multiply-add (FMA) operations. Among the new instructions introduced, the most important for the HPC applications are the set of AVX-512 instructions. They bring the ability to pack 32 double precision and 64 single precision floating point operations per second per clock cycle within the 512-bit vectors, as well as eight 64-bit and sixteen 32-bit integers, with up to two 512-bit fused-multiply add (FMA) units. In this way the width of data registers, the number of registers and the width of FMA units are doubled, compared to AVX2.

3.1.1.1.2 Cascade Lake Scalable Processors

The Cascade Lake series, not released as of early 2018, will be the successor to the Skylake SP series of processors in the server processor market. Cascade Lake falls on the optimization phase in the Intel's PAO cycle, maintaining the 14+ nm manufacturing process and architectural innovations introduced in Skylake SP, but introducing incremental improvements to the architecture. While the information on the details of these improvements are scarce, higher operating frequencies and support for DDR-T/Optane memory modules are expected. An important feature for the HPC market will be the introduction of AVX-512 Vector Neural Network Instructions (AVX-512 VNNI). Intel has released a programming reference guideline detailing these upcoming instructions in January 2018, and has supplied patches to support these instructions in the LLVM and GCC compilers.

3.1.1.1.3 Ice Lake Scalable Processors

The successor of the Skylake processor, which has been announced for 2019/20, will be the Ice Lake Xeon (ISX-SP) in 10+ nm, with up to 38 cores, 8 memory channels and up to 32 GB of High Bandwith Memory (HBM2) on board.

As successor for the Xeon Phi processor will be three ISX-H (codenamed Knights Cove), as an extended version of the Xeon SP line with 38 or 44 cores.

3.1.1.1.4 High-End Desktop, Workstation and Edge Computing Processors

While Intel did not release any new processors in the high computing power server processor segment, new processors were released in the high-end desktop and edge computing server segments.

High-end desktop and workstation segments differ from server segments, among other features, due to having a significantly lower processor core count. Edge computing servers, a relatively new concept, focus on high-speed data processing at low latency and low power requirements. Since they focus on low latency user applications instead of batch processing, they offer a relatively high core count but at a single socket configuration and a different feature set.

The Kaby Lake processor series represent the optimization phase, following the architecture phase named Skylake, retaining the 14 nm fabrication process and adding new features. The coffee Lake series, an incremental update on Kaby Lake, increased the core count from four to six.

The first processor series by Intel to be fabricated using the 10 nm fabrication process will be the Cannon Lake series. On 9th January 2018, Intel announced in a small press conference that they started the distribution of 10 nm Cannon Lake processors at the end of 2017. The initial models in these series will be targeting low power mobile computing platforms, such as ultrabooks, and will be followed by an unknown series codenamed Ice Lake.

In edge computing markets, Intel offers the Skylake-DE series and Kaby Lake-DT series, both of which are limited to 4 cores per processor.

3.1.1.2 AMD

3.1.1.2.1 EPYC

EPYC is the name of the processor product series based on the Zen microarchitecture from AMD which is aimed at the high performance server market and was released in 2017. EPYC processors are fabricated using a 14 nm fabrication process, and can be installed on cluster nodes on 1 or 2 socket configurations. Compared to the competition in the x86_64 architectures, i.e. products from Intel, they stand out due to their higher core count per socket, namely 32 cores and 64 threads per socket at higher-end products.

Processors from the EPYC product series include up to 16 MB of Level 2 cache and up to 64MB of Level 3 cache. The highest performance processor from this series, EPYC 7601, has a base frequency of 2.2 GHz, allowing operation at 2.7 GHz for all cores and at 3.2 GHz for a single core. The processor has 8 DDR4-2667 channels to the system memory, and supports a maximum memory capacity of 2 Tb per socket. The instruction set supports SSE4.2, AVX2, AES and FMA3. The reader should note that this is not an exhaustive list of supported instructions.

The current roadmap for AMD states that the Zen microarchitecture will be followed by Zen+ around mid-2018, and by Zen2 in 2019. The details for these architectures are not published, but Zen+ is expected to be fabricated at 12 nm level, and Zen2 at 7nm level.

3.1.2 ARM processors

3.1.2.1 Cavium

3.1.2.1.1 Thunder X2

Thunder X2 (see Figure 22) is an ARM SoC from Cavium that derives from the Vulcan processor, available after the Broadcom acquisition. The ThunderX2 features 54 ARM v8.2 cores produced using a 14 nm fabrication process. Compared to the previous model, Thunder X, the core count was increased by ~10% and the top frequency is slightly higher (2.6 to 3.0 GHz, up from 2.5 GHz), but the company claims 2-3x more performance.

The L1 cache size in these new SoC is 64KB for instructions and 40KB for data, while the upper level cache size is 32 MB (up from 16 MB in Thunder X). Finally, a big difference from the previous model is that Thunder X2 employs an out-of-order execution model. Due to the background of Cavium in telecommunications and networking hardware, a focus on high performance I/O is expected, and HPE claims "33% more memory bandwidth compared to the industry standard" in their product utilizing Thunder X2. Important features for this SoC are multiple PCIe3 x16 slots, 100GbE Ethernet support and 6 3.2 GHz DDR4 channels.

Figure 22. Thunder X2 high-level CPU architecture.

3.1.2.2 Qualcomm

3.1.2.2.1 Centriq 2400

Announced in November 2017, Centriq 2400 is based on the Falkor cores, the 5th custom design ARM core from Qualcomm, based on a 10 nm fabrication process. The processors support AArch64 only, and the space freed up due to the removal of 32 bit support resulted in a higher core count. The processor employs a 24 KB Level 0 cache that can be accessed without a delay in clock

cycles, in addition to a 64 KB L1 instruction cache, and 512 KB L2 cache. The Level 3 cache consists of 5 MB segments, 60 MB in total. The interconnection between cores, which also connects the cores to the L3 cache, has 256 GB/s aggregate bandwidth, which is about one quarter of the Power9 chip from IBM. Just like the Thunder X2, the chip has 6 DDR4 memory controllers, but operating at 2.67 GHz and delivering 128 GB/s aggregate peak bandwidth. However, the memory controller allows inline memory compression, potentially increasing the maximum memory bandwidth. Significantly, the chipset does not have integrated NUMA capabilities, enforcing a single socket configuration. Qualcomm claims 4x performance / price with respect to Intel Xeon 818x, and 45% better in performance /Watt.

3.1.3 POWER

3.1.3.1 IBM POWER 9

In 2017, IBM released the POWER9 series of processors, fabricated using a 14 nm process. This series come in two flavours: Scale-out and Scale-up. The Scale-out series targets the traditional, 1-2 socket cluster nodes, while the Scale-up series target 4 or more socket, high memory NUMA configurations.

The Scale-out series processors come in 2 configurations: 24 cores with 4 threads per core, and 12 cores with 8 threads per core, both providing 96 compute threads per processor. IBM stated that 24 core versions were optimized for the Linux ecosystem, whereas 12 core versions were targeting PowerVM ecosystem utilizing IBM's own operating system. Both Scale-out variants support up to 8 channels of DDR4-2667 memory, providing up to 120 GB/s of sustained bandwidth. Memory capacity is limited to 4 Tb per socket.

The Scale-up series, similarly come in 24 cores with 4 threads per core, and 12 cores with 8 threads per core, but allow 8 Tb of memory per socket and up to 230 Gb/s of sustained bandwidth.

Compared to POWER8, the key differences are the 14 nm fabrication process, a new modular architecture, NVIDIA NVLINK 2.0 support and improved single thread performance. Shorter pipelines and improved branch prediction contribute to the higher performance.

The POWER9 architecture will be followed by POWER10, which is expected to employ a 7 nm fabrication process and doubled core count (48 up from 24). POWER10 is expected to be released around 2020.

3.1.4 The Effect of Meltdown and Spectre Vulnerabilities on Performance

The recent disclosure of Meltdown and Spectre vulnerabilities have shown that speculative execution and branch prediction, as implemented in modern processors, are not side-effect free and can expose the system to side-channel attacks. Since these optimizations play important roles in the performance of the processor, the mitigation for these vulnerabilities may result in significant performance penalty.

The Meltdown vulnerability, in very simple terms, allows a process to read portions of memory that it is not privileged to access, by examining the effect of some speculatively executed code on the CPU cache. Due to the memory mode employed in modern operating systems, a process cannot access the address space of another process, as only one page table is loaded at the time. However, modern operating systems also keep the kernel portion of the page table during execution in user mode, as a means to reduce the context switch latency. Since the page table entries are present during execution in user mode, an unprivileged process may attempt to read an arbitrary location in kernel memory. As expected, this operation will fail. However, if the processor attempts to speculatively execute this instruction before raising an exception, it may load some pages into the cache, leaving a traceable side effect. All modern Intel processors speculatively execute such instructions, and are, therefore, vulnerable to this attack. AMD processors, however, stop the speculative execution if an exception will be raised and are, therefore, not vulnerable.

The initially proposed mitigation for Meltdown vulnerability in Linux was Kernel Page Table Isolation, which involved clearing the kernel page tables during context switches. This patch had a significant impact on performance, whether the processor was vulnerable or not. Later versions of the patch introduced whitelists for processor models that were not vulnerable. Currently ongoing research is on marking a portion of kernel page tables as safe, in an attempt to reduce the amount of data that need to be cleared during context switches.

The Spectre vulnerability shares the same theoretical basis, i. e. using speculative execution as a medium for side channel attacks, and has two variants. While the details of these class of vulnerabilities are beyond the scope of this report, it should be noted that unlike Meltdown, Spectre based attacks usually focus on extracting information from other user-space processes and require detailed study of the target program. However, Spectre family of attacks are found to affect a larger gamut of processors, as they exploit the branch prediction subsystem and may not result in an exception that can be detected, and require more invasive methods for mitigation.

The proposed solutions for protecting against Spectre class of attacks include user pointer sanitization in the kernel (for checking pointers passed from user-space against out-of-bounds cases, variant 1), patches to compilers such as RETPOLINE (for eliminating indirect call and jump cases for attacks employing return oriented programming techniques, variant 2), restricting timer resolution in sandboxed interpreters (for depriving the attacker from crucial high precision time measurements), and hardware level intervention such as firmware updates containing hardware counterparts of software solutions and use of PCID capability for selective cache flushing.

The reason for providing this extensive (yet not exhaustive) list of mitigations is to emphasize that these vulnerabilities have led to the implementation of various sanity checks, all of which come at varying degrees of performance penalties. Furthermore, since many of these mitigations require operating system level patches including kernel updates, in addition to repackaging of existing software, they pose a serious challenge to the HPC ecosystem that has historically stood on the conservative side on the subject of updates.

3.2 Highly parallel components/compute engines

3.2.1 FPGA: Intel Stratix 10

Following the acquisition of Altera, Intel incorporated FPGAs as a critical part of its own growth strategy in the data centre market [28]. The first product to target a much broader customer base than the already established FPGA market is the Stratix 10, the latest product in the Intel-Altera Stratix line. Intel claims that the acquisition provided the former Altera technology development an “Intel advantage” in the form of integration expertise and industrial capabilities. This eventually led to the first FPGA product able to replace the traditional 2D MOSFET transistors, a common foundation for the current FPGA market segment, with a 3D FinFET (Tri-Gate) transistor technology integrated at 14 nm, the first time such a scale is employed in an FPGA product [29].

The new Intel HyperFlex architecture boasts an ARM Cortex A53 as a functional block connected to the programmable FPGA network, which in turn sees systolic registers placed at each routing node (a “registers everywhere” approach). Intel claims this is an enabling factor for a streamlined development and performance tuning experience.

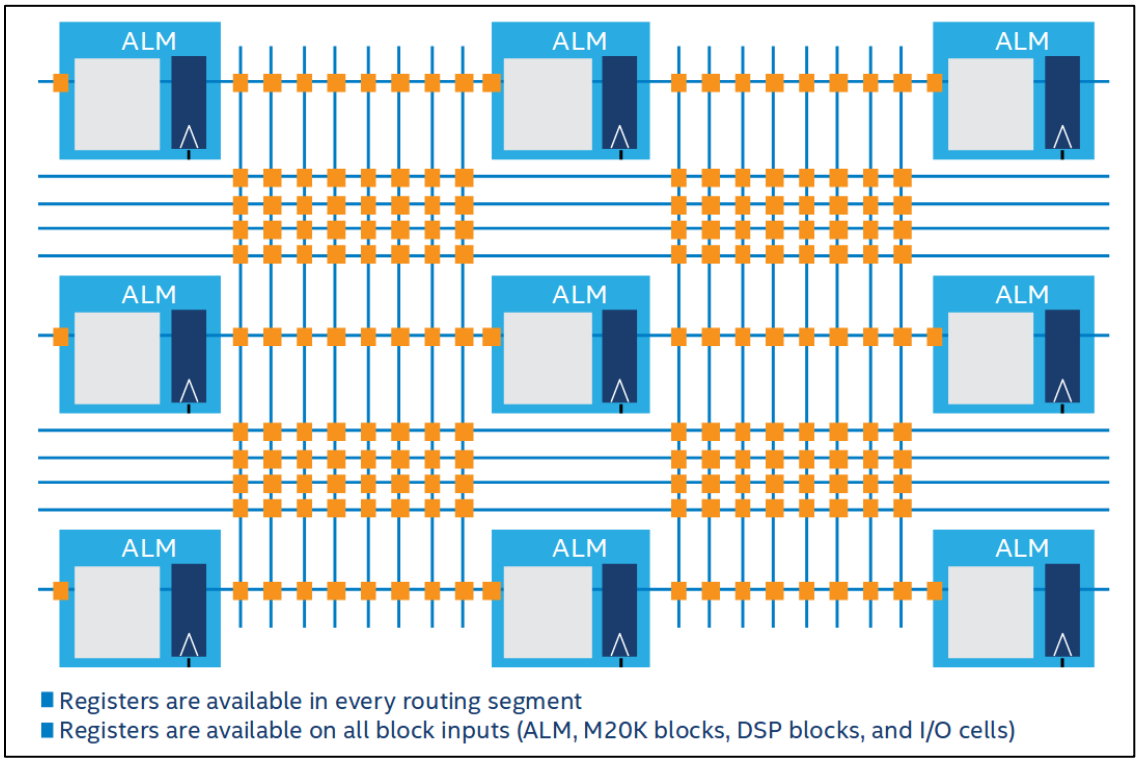

Figure 23. Intel HyperFlex architecture, "registers everywhere" approach [30]

The key feature that differentiates the HyperFlex from traditional FPGA architectures is the presence of dedicated, single-precision IEEE 754 “hard” (opposed to “soft”, or programmable) units, an adder and a multiplier, available at each Digital Signal Processor (DSP) site. While the usual DSP units allow the designer to implement variable precision, fixed point custom operations, the floating point units allow to deal with floating point data flows without the expense of additional programmable logic. Furthermore, these IEEE 754 units can be organized in column-shaped

blocks, capable of supporting typical linear algebra functions as well as more traditional (in the established FPGA market) signal processing functions like highly parallel FFTs and impulse response filters. Intel claims that, given a combined and efficient use of floating and fixed point DSPs, the top-of-the-line 2.8 MLE Stratix 10 is capable of 9.3 TFlops, assisted by 1 Tb/s of memory bandwidth (thanks to the HBM2 system-in-package integration) in a power efficient package. This leads to an estimated 80 GFlops/W theoretical efficiency [31]. The product card is equipped with standard interfaces like a PCIe Gen2 x8, a DDR3 channel interface and an on-board 10G Ethernet controller.

While the traditional FPGA development workflow involves steps that are familiar to hardware designers (like clock distribution and finding the right balance between clock frequency and number of programmed units), the Stratix 10 ships with a full OpenCL 2.0 compliant stack chosen as the paradigm in charge of abstracting away the traditional FPGA hardware-based development workflow. The OpenCL paradigm allows the developer to write high-level code and have the Altera Quartus Prime compiler generate custom hardware for each accelerated instruction [31]. However, even Intel admits that, despite the presence of a standard interface, the way an FPGA is exploited is substantially different from other massively parallel hardware. It is worth noting though that the Stratix SDK ships with an emulator capable of producing detailed optimization reports and this can be a cornerstone of the iterative development workflow.

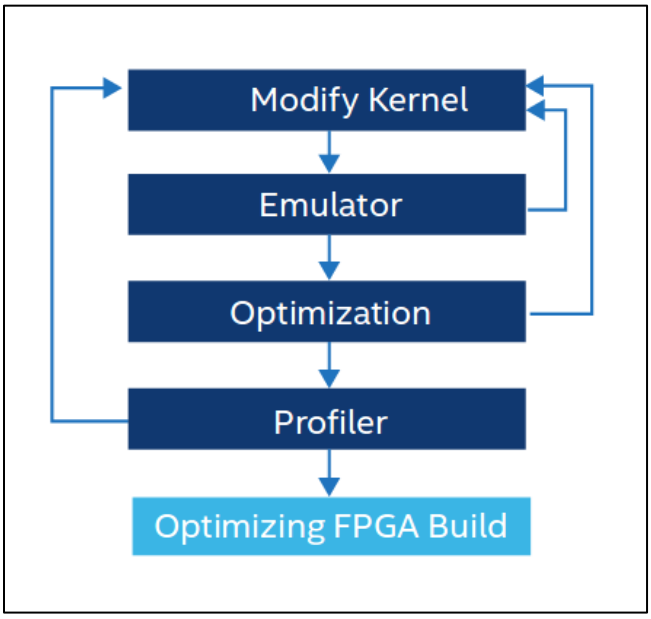

Figure 24. FPGA development workflow [32]

While still targeted at the FPGA traditional applications, Stratix 10 Intel is steering its Altera assets towards a much broader market. This places the Stratix 10 as a “natural competition” for GP-GPUs in cloud and HPC data centres [31].

3.2.2 Manycore: PEZY

Occupying the first three positions in the Green500 Nov. 2017 list, the PEZY (which stands for Peta, Exa, Zetta, Yotta) manycore is the accelerator chip that boosts Japan's most power efficient HPC installations.

Powering the ZettaScaler-2.2 architecture, currently at the first Green500 position with the Shoubu system B, is the PEZY-SC2. The SC2 is a second-generation chip, manufactured at 16 nm integration scale, featuring 2048 MIPS cores (nicknamed "cities") with 8-way SMT and 128 bit SIMD units each, totalling over 16k hardware threads on a single chip. Operating at 1 GHz with 4 flops per cycle per core, the SC2 has a peak performance of 8.2 TFlops at IEEE 754 single-precision [33]. The Shoubu system B, integrated with immersion-cooling cabinets, topped the list with a 17 GFLOPS/W efficiency peak. In order to leverage the massively parallel PEZY systems, a custom designed subset of OpenCL, named PZCL, is provided and recommended as the "official" paradigm [34].

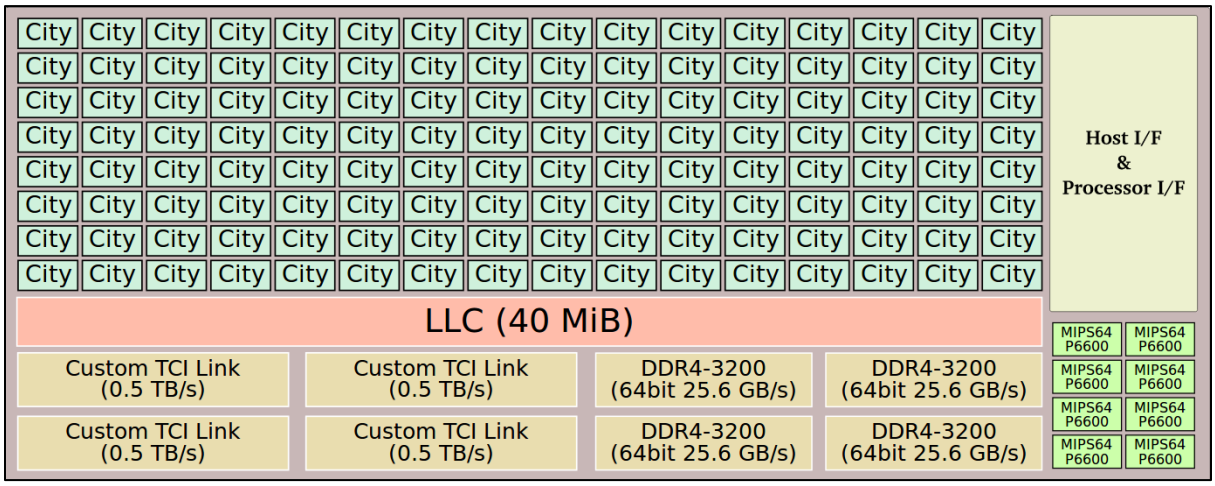

Figure 25. PEZY-SC2 main block architecture [35]

A key aspect is the Japanese-only nature of the HPC initiative: the ZettaScaler system series was born from a joint effort of NEDO governmental funding program and three private companies. These are the following, each one being founded and managed by the ExaScaler CEO, Dr. Motoaki Saito [36]:

- PEZY Computing Co. Ltd., a fabless semiconductor company which develops the manycore accelerator (that are actually produced by Taiwan Semiconductor Manufacturing Company);
- ExaScaler Inc., focused on highly-efficient, submersion liquid-cooling;
- Ultra Memory Inc., which develops the 3D multi-layer memory system.

As additional evidence of Japan's commitment, PEZY Computing has already laid out a roadmap extending up into the 2020s that foresees breaking the 1 EFlops mark with the ZettaScaler-3.0 supercomputer in late 2019 based on the upcoming PEZY-SC3 development iteration. This is expected to yield a 40 GFlops/W of peak efficiency at 7 nm integration scale. It is notable to mention that both the PEZY-SC3 and PEZY-SC4 are expected to replace the standard PCIe

controllers with silicon photonics (likely optical PCIe) [33]. The same development roadmap extends even further, planning to reach 60 GFlops/W in 2020 with the fourth generation PEZY-SC4 at 5 nm integration scale.

3.2.3 Open source: RISC-V

Being the first open-source ISA targeting real world processing beyond teaching purposes, RISC-V took off and gathered the efforts of many contributors both from academia and industry [37]. In contrast to most ISAs, the RISC-V ISA can be freely used for any purpose, permitting anyone to design, manufacture and sell RISC-V chips and software. Even though it is not the first open ISA, it is significant because it is designed to be useful in modern computerized devices such as warehouse-scale cloud computers, high-end mobile phones and the smallest embedded systems. The instruction set also has a substantial body of supporting software. The design pillars on which RISC-V is built are:

- small, carefully designed and easy to implement ISA, flexible enough to avoid bloating during its own evolution;
- low power;
- modular and customizable, with plenty of optional extensions (like various IEEE754 precisions, compressed instructions, atomic instructions and transactional memory among many others) that allow designers to craft their own chip to fit their specific needs.

The architectural specification also provides extensions for SIMD units, both packed and variable length. While the former is meant for extreme low-power, low budget applications, the latter is notably based on a novel, flexible design. This flexible SIMD unit is a general-purpose, mixed-precision vector processor with vector length selectable at runtime at instruction level. This allows for easy code porting to CPUs with different vector lengths, ideally without recompiling [38] as opposed to the widespread short-vector SIMD approach where applications need to be at least recompiled at each evolution step (e.g. from AVX256 to AVX512).

The architecture specifications and ISA are freely available [38] and several open source implementations in synthesis languages (mainly System Verilog), publicly available on GitHub, are attracting several contributors from both industry and academia. Notable among ongoing academic efforts include:

- the Parallel Ultra-Low Power (PULP) Platform, designed for energy-efficient IoT computing. As a part of PULP, ETH Zurich and the University of Bologna have cooperatively developed the open-source, parallel, in-order PULPino processor targeted at scientific computing. Inside the same framework, several ongoing research efforts are focused on providing the developer with a fully- fledged, efficient OpenMP implementation on top of the PULP platform [39];
- the 64-bit Rocket Chip from Berkeley may suit compact, low-power intermediate computers such as personal devices;
- the 64-bit Berkeley Out of Order Machine (BOOM) utilizes much of the infrastructure created for the Rocket Chip, and may be usable for personal, supercomputer and warehouse-scale computers;
- the Open Transprecision Computing project (OPRECOMP), in the framework of the EU-funded Horizon 2020 initiative, aims at leveraging RISC-V flexibility to explore the field

of variable precision algorithms and hardware. In fact, one aim of the project is to build a transprecision capable system based on the PULP processor coupled to an IBM Power8/9+GPU cluster (IBM, ETH Zurich and the University of Bologna are all partners of OPRECOMP).

- the European Processor Initiative aims at building an Exascale system based on EU-developed technology on top of the RISC-V architecture [40].

Industrial efforts include:

- Adapteva is basing its own next generation of many-core accelerators on RISC-V architecture, pointing out that one of its selling points is design simplicity and neatness: "The RISC-V team had the luxury of learning from mistakes made over the last 50 years of computer architecture development and has left all of the heavy baggage behind" [41];
- NVIDIA plans to use RISC-V chips in its GeForce graphics cards [42];
- Western Digital announced its plans to equip all of its hardware products with RISC-V controllers due to the efficient power design on which the architecture is built [43];
- The US-based SiFive introduced the first Linux capable SOC based on their own (open source) chip implementation, the HiFive [44]. Although still in an early stage, the availability of development platforms is a key enabling factor to allow the architecture to become much more widespread.

As an open and collaborative ecosystem, the RISC-V architecture is currently gaining traction in a much broader market and the RISC-V Foundation boasts among its supporters major hardware vendors like AMD, IBM, NVIDIA, Qualcomm, Mellanox and Western Digital as well as prominent players in the computing industry like Google and Microsoft. Furthermore, the extreme diversity of applications is another evidence of RISC-V degree of flexibility and customization that allows uses ranging from ultra-low-power IoT to novel HPC accelerators.

3.2.4 GP-GPU: NVIDIA Volta

Announced in May 2017, the latest NVIDIA microarchitecture, the Volta, sees its first product incarnation with the Tesla V100 GPU card. Given its 15 GFlops/W efficiency peak, the 12 nm integrated V100 occupies the 4th position in the Green500 November 2017 list with a NVIDIA-operated DGX Saturn Volta system. Nonetheless, although its outstanding performance this system has been surpassed by three PEZY-based, Japanese systems by a significant factor (the first position breaks the 17 GFlops/W mark).

New features introduced by the Volta architecture focus on three key aspects:

1. Programming experience: CUDA 9 introduces cooperative groups into the CUDA language, a new programming model aimed at a streamlined programming experience when blocks of CUDA threads must be synchronized;
2. Increased bandwidth: High Bandwidth Memory 2 (HBM2) and NVLink 2.0 (estimated to provide 25 Gb/s per lane as opposed to the 2 Gb/s PCIe 3.0);
3. Neural network applications performance: dedicated Tensor Cores are integrated into the die.

The most notable strategic factor introduced by the Volta architecture is that, for the first time, a specialized hardware, the dedicated neural-network processor Tensor Core, makes its appearance

in the consumer market on board of an off-the-shelf product. A Tensor Core is a highly specialized unit that multiplies two 4×4 FP16 (IEEE754 half-precision) matrices and then adds a third FP16 (optionally IEEE754 single-precision, or FP32 in the NVIDIA jargon) matrix to the result via fused multiply-add operations. This is significant as this kind of operation lies at the heart of neural networks training.

While both OpenCL and OpenACC open standards are supported, the privileged way to leverage Volta hardware remains NVIDIA's own CUDA language. Even if NVIDIA's current focus is clearly on the market where neural network applications are booming, the CUDA ecosystem provides a vast array of traditional, multidisciplinary HPC applications optimized out-of-the-box for the Volta architecture [45].

3.3 Memory and storage technologies

For the last two years, the HPC industry has been preparing users for a new era of memory hierarchy, where the flat model using just one type of memory – DRAM will be replaced with a combination of:

- very high bandwidth memory (HBM or HMC) with small capacity compared to DRAM,
- slower but persistent memory (NVM) with bigger capacity than DRAM,
- standard memory to fill the gap between these two extremes (both for capacity and bandwidth).

3.3.1 HBM, HMC and GDDR

The second generation of HBM is available only on GPU accelerators from NVIDIA and AMD. The current generation (KNL) of the Intel Xeon Phi products offers a very similar technology called HMC. Both can be seen as 3D memory, differing only by the consortium that defines the standards [46]. Although HBM has a much higher bandwidth than DRAM, it has a slightly higher latency because of the stacking of silicon layers, reaching up to 18% higher latency compared to DDR4 DRAM [47].

	Total Capacity	**Total Bandwidth**
NVIDIA HBM2	32 GB (4 stacks)	900 GB/s
AMD HBM2	16 GB (2 stacks)	484 GB/s (ECC off)
Intel MCDRAM	16 GB	400 GB/s
Samsung Aquabolt	32 GB (4 stacks)	1228.8 GB/s (theoretical)

Table 7. HBM memory applications overview

Recently, Samsung announced the start of mass production of a modernized HBM2 (called Aquabolt) which will bring an increase of the bandwidth by using higher frequency stacks from 2 Gb/s (defined by the JEDEC) to 2.4 Gb/s delivering 1228.8GB/s and 32GB in total with stack size of 8GB [48].

GDDR is a competitive technology to HBM/HMC especially in the field of GPUs. Currently, both SK Hynix and Samsung announced mass production of GDDR6. At GTC2018, NVIDIA

announced to present GDDR6 from Hynix on majority of their future GPUs [49]. An 8Gb chip will run at 14 Gb/s per pin giving 56 GB/s of bandwidth.

3.3.2 DRAM

The current standard of DDR4 memories is 2666 MT/s per DIMM support by the majority of CPU vendors including latest Intel Skylake-SP, AMD EPYC and Cavium ThunderX2. These CPUs differ in memory channels per chip configuration with 6 channels for all the Intel CPUs and 8 channels per chip for the high-end AMD and Cavium, delivering a theoretical maximal bandwidth of 127GB/s on Intel or 170Gb/s on AMD/Cavium per socket. By the end of 2018, a new generation of Intel CPUs named Cascade Lake will be available with memory controllers supporting 2933 MT/s DDR4 DIMMs, still providing 6 channels per socket. It should be mentioned that JEDEC standard defines up to 3200 MT/s DDR4 technology, so other vendors than Intel might go to even higher bandwidth rates on this technology.

3.3.3 NVM

Non-volatile memory (NVM) is a new class of storage class memory that is byte addressable and persistent so, unlike DRAM, doesn't lose state on power loss and can be accessed using memory instructions or memory APIs even after the end of the process that created or last modified them. While NVM boasts similar density and near DRAM speeds, it offers a lower cost, larger capacity storage layer that traditional memory (byte addressable, accessible via DMA) as well as a lower latency, higher durability persistent storage.

Nowadays, the data centre copes with the increasing need of low latency storage using a combined approach: on the storage server, a thin layer of DRAM caches fronts a huge amount of high-capacity, persistent SSDs based on traditional NAND technology. With the introduction of NVMs, a new intermediate layer between DRAM and NAND enters the storage hierarchy that, for the first time, acts as a fast, inexpensive and persistent memory layer that can serve as system memory and storage at the same time [50]. This layer allows system architects to adopt "disaggregation": since NVM DIMMs can be accessed through RDMA, bypassing the OS I/O stack to minimize latency, it acts as an enabling factor in the effort of pulling the mass storage out of deep centralized pools and spreading it across the data centre, as close as possible to the computing elements. This allows applications to efficiently keep entire datasets resident in memory, providing to memory-centric paradigms, such as Spark, a natural opportunity to perform operations in the DIMMs, rather than in the CPUs.

The first NVM product available is the Intel Optane range, which is based on the Intel 3D XPoint technology that Intel claims to have ~8x to 10x greater density than usual DRAM due to its stackable nature and due to the fact that, unlike traditional DRAM, each memory cell doesn't need a helper transistor. This leads to much more compact, power efficient components. Moreover, Intel claims that this kind of technology is not significantly impacted by the number of write cycles – thus making it more durable than traditional solid state storage [51].

NVM can be accessed both as a standard file system and a standard raw (character) device, due to the support already shipping in all major OS, allowing applications to leverage its advantages

without the need to change a single line of code. However, the most efficient access is performed via Direct Access (DAX) calls that enable direct load/store operations on files stored in persistent memory, bypassing the OS I/O stack. This requires the code to perform proper API calls, but allows the developer to leverage all the advantages of the new storage layer.

The DAX API has been integrated in regular file system drivers (currently EXT4 and ZFS on Linux and NTFS on Windows) making the file system "persistent-aware": when an application opens a memory mapped file on this file system, it has direct access to the persistent region while, without DAX support, the page cache is generally used to buffer reads and writes to files, and requires an extra copy operation.

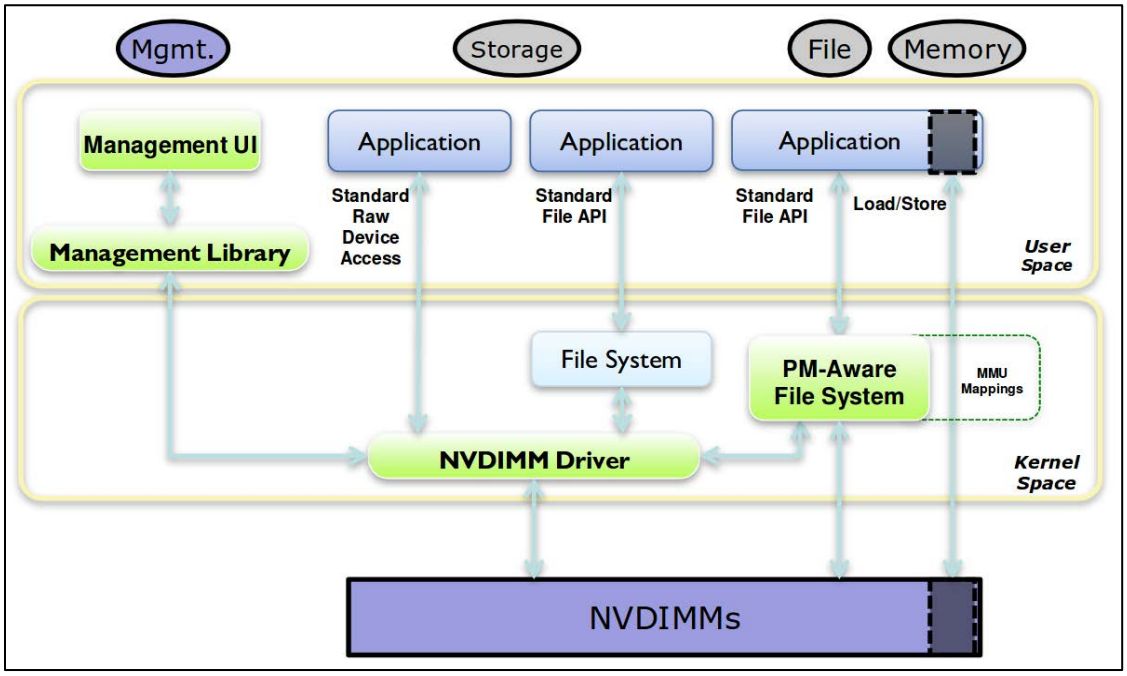

Figure 26. Persistent memory programming model depicting all kinds of NVM access (from left to right): raw device, file system, PM-aware file system and DAX [52]

The Storage Networking Industry Association (SNIA) specification defines recommended behaviour between various user space and operating system kernel components supporting NVM. This specification does not describe a specific API. Instead, the intent is to enable common NVM behaviour to be exposed by multiple operating system-specific interfaces. Some of the techniques used in this model include memory mapped files and direct access (DAX). The same association provides a Persistent Memory Developer Kit, or PMDK (formerly NVML): a platform neutral and vendor neutral collection of open source APIs enabling the developer to access NVM [51]. On top of low level APIs, PMDK ships an open source collection of libraries providing: (a) convenient and performance-tuned APIs for NVM like transactional access, (b) persistent memory allocators for several languages (`std::allocator` compliant for C++, for example) and (c) transactional object stores. There is work in progress for adding features like remote persistent memory via RDMA.

NVM provides growing data centres the ability to run both in memory and traditional applications on a single infrastructure acting as the convergence of storage and memory, which for decades were two separate computing domains.

3.3.4 Tapes

A streaming media like tape will never be able to match solid-state memories, but it can be faster at streaming data than hard drives once it has found the right position. Thus, it needs to be used in a proper way and not placed in the hot data path for computations.

Tape technology is traditionally placed very far down the memory and storage hierarchy, and is usually tiered behind disk storage. In recent years, attempts have been made to skip the disk part, or at least replace spinning disks with flash storage. This combination has been dubbed FLAPE (from the words FLash and tAPE), with the idea being that smaller flash storage can handle random access requests for frequently used data and metadata with tape providing bulk storage of data. The market here is more big data and archiving than compute. So far, this has not seen much publicized deployment and is highly dependent on data management software to support this use case.

Chapter 5.2 contains more on tape technology.

3.4 Interconnect

Interconnects used in the HPC world today can be distinguished into a few main categories.

The main category could be called “HPC computing” where the low latency and high bandwidth are dominant. The technologies most represented in this category are Infiniband (mostly delivered by Mellanox) with 32.6%, Cray Aries with 8.2% and Intel Omni-Path with 6.8%. Other representatives are custom technologies specially developed and used in few big systems, like the Sunway in TaihuLight, TOFU-2 in K-Computer and TH Express-2 in Tianhe-2 machines. Older custom technologies not available anymore in the market such as Cray Gemini or IBM BlueGene/Q and are slowly disappearing. Last year, two systems with the new custom BXI interconnect have made it to the list and demonstrating that this might be another option for HPC clusters developed by BULL.

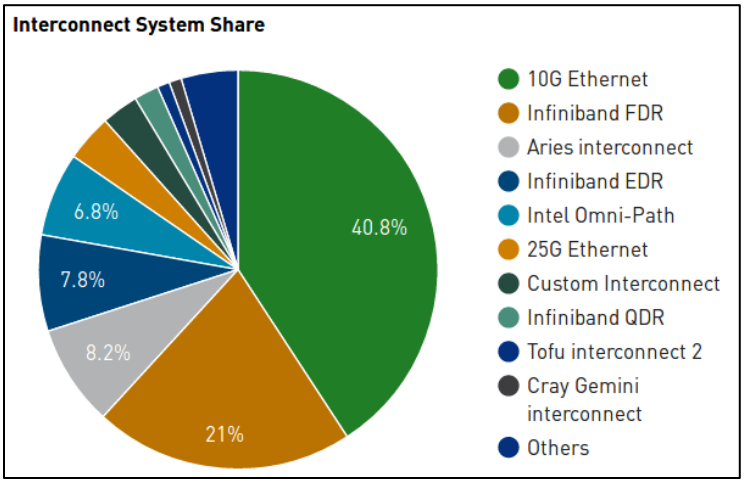

Figure 27. TOP500 list (November 2017) Interconnect market share by systems

The next category could be called "HTC/cloud/hyperscaler computing" where bandwidth matters too, but there is less pressure on the latency. The typical technology for this category is Ethernet and in the current TOP500 edition, the 10 Gigabit Ethernet technology represents 40.8% of listed systems. These are mostly Chinese not originally designed for HPC workload but benchmarked with HPL. Also, the newer 25 Gigabit and 100 Gigabit Ethernet starts to show up in the list, but the main market for this technology lies in the big hyperscaler datacentres like Google, Facebook, Microsoft, Amazon, Baidu, Alibaba, Tencent and China Mobile.

The last category of interconnects, interesting somehow to the HPC world, could be called "internal interconnects" with representatives connecting only few elements (CPU, GPU, memory, NIC, etc.) like Intel UPI, AMD Infinity Fabric, NVIDIA NVLink, CAPI/OpenCAPI, but also with representatives connecting more elements like HPE NUMAlink 8 and the still to be developed GenZ.

3.4.1 Omni-Path, Infiniband, Aries, BXI

There is no public news about when the second generation of Omni-Path will be available, which should address 200 Gb/s link speed and thus match the Infiniband HDR rival from Mellanox. The current highlights of Omni-Path gen 1 can be summarized into the TOR switch portfolio with up to 48 ports at 100Gb/s with port-to-port latency around 110 ns, director class switch with up to 768 ports and server cards with one 100Gb/s port and approximately 160M messages/s rate. This is based on the 25 Gb/s SERDESes.

Mellanox with its Quantum line of Infiniband HDR products also didn't reveal anything new since last year and so the main characteristics of the products are:

- Top of rack switch QM8700 with 40 HDR ports running at 200Gb/s or using split cables 80 ports at 100Gb/s with port-to-port latency around 90ns.
- Director level switch CS8500 with up to 800 ports at 200Gb/s or again 1600 ports at 100Gb/s with split cables.
- ConnectX-6 server cards with up to 2 ports at 200 Gb/s and up to 200M messages/s rate.

The HDR generation of Infiniband relies on the 50 Gb/s SERDESes where 4 of them are bonded to form the 200 Gb/s link. Since Mellanox has already for their Ethernet line the option to bond 8 of such links, 400 Gb/s on Infiniband (the NDR generation) is to be expected, although no official public information is known when this will happen.

The Aries interconnect is exclusive to the XC line of supercomputers from Cray and doesn't have any updates in specifications. Only a strong message for big data and machine learning effectiveness was given by Cray during last year. Trying to compare the key aspects of the technology with Infiniband and Omni-path would lead to something like the TOR switch with 40 ports to connect other switches at 37.5-42 Gb/s and 4 ports to compute nodes at 81.6Gb/s. The server cards offer the 81.6 Gb/s. The SERDESes used for this operate at 12.5 Gb/s for optical switch connections, 14Gb/s for electrical switch connections and 8Gb/s for the server cards connections. The message rate for this reaches up to 120M messages/s.

The BXI is still considered as a BULL-only option, maybe even only for Sequana X1000 line of their supercomputers. Technically it presents the 48 ports at 100 Gb/s and "less than 1 microsecond" latency on the switch level, server card with one 100 Gb/s port and message rates from 110M (unidirectional) messages/s to 160M (bi-directional) messages/s.

3.4.2 Ethernet

There has been much progress in Ethernet technology driven by the hyperscalers, which bring possible usage of the technology for HPC as well. Utilizing 8 links of the new 50 Gb/s SERDESes, adopting the PAM-4 signalling already used in Infiniband, Ethernet now shines with 400 Gb/s links. However, it still has 3-4 times higher latencies compared to Infiniband, where the best ones on the port-to-port switches from Mellanox give 300ns and approximately 420ns on Broadcom switches.

The top of line Broadcom Tomahawk 3 based switch will have up to 32 ports at 400 Gb/s or 64 ports at 200 Gb/s or 128 ports at 100 Gb/s supporting both the new PAM-4 signalling or the legacy NRZ especially for long reach optics.

Mellanox has with its Spectrum-2 chips the best switch configuration at 16 400 Gb/s ports or 32 200 Gb/s or 64 100 Gb/s ports or 128 10/25/50Gb/s ports.

In both cases, these products are officially advertised on the web sites of the companies, but not yet buyable, with the general availability expected to be during 2018. Broadcom technology might be sold by Arista, which seems to have the biggest volumes of sales in this market.

3.4.3 Numalink, GenZ

The evolution of former SGI NUMAlink (now as part of HPE Superdome FLEX line of servers) has arrived at the 8th generation. The new generation changed the signalling from 50 Gb/s (SGI UV2000/3000) or 56 Gb/s (SGI UV300) to 100 Gb/s and reduced the latency. Compared to the previous generation of servers (HPE MC990X or SGI UV300) in a 32 sockets system the latency has gone down by approximately 25% (365ns compared to previous 480ns) and bandwidth is almost 50% higher [53]. This is aligned with the new Intel UPI links on the latest Skylake CPUs

which also increased bandwidth from 9.6 GT/s to 10.4 GT/s. Currently the 32 sockets system supporting up to 896 cores and 48Tb RAM is the maximum you can order from HPE (aligned for one full rack), although the interconnect is designed to support up to 128 sockets. The previous generation (NUMAlink 6) supported up to 256 sockets, but as the DIMM sizes as well as the expected NVDIMM sizes grow it seems that 128 sockets will be more than enough for the market. The potential in 4 racks using Intel 3DXpoint NVDIMMs would support in total 480Tb of memory (384 Tb in NVRAM and additional 96TB normal RAM). This new generation of NUMAlink aims not only for really big SMP machines, but looks also at the smaller ones supporting the cheaper Gold line of Skylake CPUs (61xx) and allowing to build up to a 32 socket system, all of them directly connected with just one NUMAlink hop between them, taking advantage of the fact, that one NUMAlink ASIC is actually 2 NICs and a switch in a single piece of silicon. Another option is to use adaptive routing in the NUMAlink network and have point-to-point traffic over multiple links between the sockets and so to create a "multi-rail cluster" with higher bandwidth between the sockets.

A yet to be done interconnect for internal but also for rack-level connections was proposed by the GenZ consortium. It has around 50 members with chip makers like AMD, IBM, Samsung, Broadcom, Arm, Cavium and Micron, system makers like HPE, Lenovo, Cray and Dell, interconnect makers like Mellanox and Numascale and storage makers like Seagate, Toshiba, Western Digital and SK Hynix. One of the aims is to enable high-speed interconnects between components like CPUs, memory (both RAM and different NVMs), GPUs and FPGAs, but also to enable high bandwidth, low latency interconnect between systems. One way is to act as a PCIe replacement with much higher capacity and it is proposed to have 56 GT/s in the beginning with a path to 112GT/s and more. For comparison, the PCIe gen 4 has 16 GT/s and the NVLink gen 2.0 has 25GT/s. But GenZ is not only another proposal for PCIe replacement like CCIX or OpenCAPI. It also aims to support direct messaging meaning turning GenZ into something like Infiniband or Ethernet providing direct support for MPI, SHMEM and sockets using OpenFabric's OFI Library. The specification of GenZ version 1.0 as an open standard is already available on the web site with products expected in 2019. It will be interesting to see, if this approach will succeed and replace the PCIe and somehow cut the portion of the market of traditional network technologies.

3.4.4 BlueGene2, EXTOLL, TH Express-2, TOFU-2

There was no progress in the last two years for these interconnects, so the description provided by the previous report is still valid [54].

4 Overview of vendor solutions

4.1 Atos

Bull systems are under ATOS technologies HPC branch. Atos/Bull currently has 17 supercomputers ranked in the TOP500 [1].

Sequana X1000

Sequana X1000 is the multi-petascale line for completely Direct Liquid Cooled (DLC) systems from BULL promising PUE “close to” 1.0 and up to 40°C inlet temperature. The system consists of two compute cabinets and an interconnect cabinet in between. Single compute cabinet holds up to 144 compute nodes and a hydraulic module to cool them and the associated liquid cooled power modules at the top of the cabinet. The switch cabinet contains level 1 DLC switches - BXI or Infiniband EDR, level 2 Direct Liquid Cooled switches (BXI or EDR).

In the Bull Sequana X1000 the computing resources are grouped into cells. Each cell tightly integrates compute nodes, interconnect switches, redundant power supply units, redundant liquid cooling heat exchangers, distributed management and diskless support. Each cell can, therefore, contain up to 288 dual-socket Intel Xeon nodes OR, 288 single-socket Intel Xeon Phi nodes or 96 dual-socket Intel Xeon nodes with 4 NVIDIA Pascal GPUs.

Bull Sequana has blades available for several Intel® Xeon processors (Broadwell, Skylake), Xeon Phi processors (KNL) and also has announced blade for ARM processors.

Sequana X800

The X800 line is focused on memory computing and big data analytics mainly. The maximum single instance OS system can reach up to 32 sockets (896 cores) and 384 DIMMs (48TB of RAM) with up to 80 PCIe slots for GPU accelerators, NVMe storage and other components for fast processing.

Sequana X550

Smaller HPC clusters (up to hundreds of nodes) especially when air cooling is the only option can be built on the X550 line which is a blade system with chassis supporting up to 20 two-socket nodes with integrated IB EDR or OPA switch and support for GPU accelerators. This modernized version of the previous generation of B510 and B515 support also NVMe storage.

4.2 Cray

Cray’s products have been divided between Compute, Storage and Analytics. Compute product line has both XC series supercomputers and CS series cluster supercomputers. The storage product line includes Cray ClusterStor Lustre solutions acquired from Seagate and also Cray Datawarp I/O accelerators. For the analytics products line Cray has Urika analytics software environments.

Of current TOP500 [1] list Cray has 53 systems.

Cray's XC product line is for supercomputers. The XC series integrates a combination of vertical liquid coil units per compute cabinet and transverse air flow reused through the system. The newest product Cray XC50 supercomputer supports the newest generation of CPU and GPU processors, NVIDIA Tesla P100 PCIe GPUs and Intel Xeon Scalable processors coupled with the Aries network and high-performance software environment. The Cray XC50 compute blade implements two Intel Xeon processors per compute node and four compute nodes per blade. Compute blades stack 16 to a chassis and each cabinet can be populated with up to three chassis, resulting in 384 sockets per cabinet and providing a performance of more than 619 TFlops per cabinet. Cray XC50 supercomputers can be configured up to hundreds of cabinets and upgraded to nearly 300 PFlops per system with CPU blades and over 500 PFlops per system with a combination of CPU and GPU blades.

The Cray XC50-AC air-cooled supercomputer, supporting NVIDIA Tesla P100 PCIe GPUs and Intel Xeon Scalable processors, delivers up to 236 TFlops peak performance in a 24" cabinet with no requirement for liquid cooling or extra blower cabinets. Ideal for dedicated test, development, AI and analytics use cases, the air-cooled XC50 system offers all of the benefits of our XC50 supercomputer in a smaller form factor.

Both XC50 and XC50-AC lines newly support ARM-based Cavium ThunderX2 processors. The ARM option has the same level of technology support including the Aries interconnect, Cray Linux Environment and Cray Programming Environment meaning that the end user gets a complete suite of compiler, libraries and development tools to work on this platform the same way as on X86. First evaluations of this platform were done by end users from GW4 Alliance and Met Office in the UK who will have the first system called Isambard sold this year.

The Cray DataWarp I/O acceleration option for the XC series supercomputer utilizes flash storage to speed up storage performance to applications and compute nodes in a variety of scenarios.

Cray Urika-XC analytics software suite was launched for Cray XC supercomputers. With the Cray Urika-XC software suite, analytics and Artificial Intelligence (AI) workloads can run alongside scientific modelling and simulations on Cray XC supercomputers, eliminating costly and time-consuming movement of data between systems.

Cray provides also cluster systems in CS product line. Cray CS500 system supports for 64-bit Intel Xeon Scalable processors Optional support for Intel Xeon Phi processors and NVIDIA Tesla GPU computing accelerators. FDR or EDR InfiniBand with Connect-IB, Intel Omni-Path Host Fabric Interface. Air cooled, up to 72 nodes per rack cabinet. CS500 system compute environment can scale to over 11,000 compute nodes and 40 PFlops peak.

The Cray CS400-AC is an air-cooled cluster supercomputer, highly scalable and modular platform based on the latest x86 processing, co-processing and accelerator technologies from Intel and NVIDIA. The Cray CS400-AC high-performance compute environment capable of scaling to over 27,000 compute nodes and 46 peak PFlops.

The CS400-LC system is direct-to-chip warm water-cooled cluster supercomputer. Designed for significant energy savings, it features liquid-cooling technology that uses heat exchangers instead of chillers to cool system components. A single high-density rack dedicated to GPU computation

can deliver up to 658 TFlops of double-precision performance. For machine learning, where integer operations matter, a single CS-Storm 500GX server node can deliver up to 170 Tops (Tera operations per second).

The Cray CS-Storm 500GT configuration scales up to ten NVIDIA Tesla Pascal P40 or P100 GPUs or Nallatech FPGAs. A single high-density rack dedicated to GPU computation can deliver up to 658 TFlops of double-precision performance. For machine learning, where integer operations matter, a single CS-Storm 500GX server node can deliver up to 170 TOPS (tera operations per second).

The Cray CS-Storm 500NX configuration scales up to eight NVIDIA Tesla Pascal P100 SXM2 GPUs using NVIDIA NVLink to reduce latency and increase bandwidth between GPU-to-GPU communications, enabling larger models and faster results for AI and deep learning neural network training.

4.3 Dell EMC

Dell EMC provides HPC products with its PowerEdge C series product line.

The C series has both 2U and 1U servers of dense, performance optimized compute nodes for scale-out workloads.

PowerEdge C6410 provides up to 4 independent hot-swappable 2-socket compute nodes in a 2U C6400 chassis and PowerEdge C6320p has up to 4 independent hot-swappable a-socket nodes in C6300 chassis. PowerEdge C4130 is a 1U server with 2 CPU sockets and up to 4 GPU sockets of 300W.

Dell co-operates with CoolIT to provide liquid-cooled solutions for data centres.

PowerEdge C6420 has 4 nodes in 2U form factor with CoolIT Systems rack based Direct Contact Liquid Cooling (DCLC) technology to support higher wattage processors. Each 1U half-wide compute sled (1 node) includes dedicated liquid cooling to high wattage dual processors. The cold plate solution designed and manufactured by CoolIT uses room-temperature water to cool the CPUs, eliminating the need for chilled water and lowers overall energy costs by 56%.

CoolIT rack DCLC utilizes a three-module building block approach.

CoolIT Systems Server module has cold-plates, specifically designed for use with Intel Xeon SP processors are passive CPU cooling solutions managed via a centralized pumping architecture. These cold plate assemblies replace heatsinks.

In Manifold Module coolant tubes come out of each sled and connect to a manifold unit. Made of reliable stainless steel and 100% non-drip quick disconnects, Rack Manifolds for PowerEdge C6420 are installed at the back of the rack.

CoolIT Systems Rack DCLC product line offers a variety of Heat Exchange Modules depending on load requirements and availability of facility water.

4.4 HPE

The HPE product range for HPC systems is wide and consists of both liquid and air cooled computer systems from 1U servers to PFlops scale supercomputers. HPE currently has over 100 supercomputers ranked in the TOP500 (Nov2017)

The HPE SGI 8600 System is a liquid cooled, tray-based, scalable, high-density clustered computer system designed for HPC workloads at PFlops speeds. Compute nodes are available with the Intel® Xeon® Processor Scalable Family, the Intel Xeon Phi processor, or the Intel Xeon Processor Scalable Family with NVIDIA Tesla SXM2 GPUs.

The HPE SGI 8600 System is based on a compact E-Cell design with up to 36 trays, 144 nodes, and 288 Intel Xeon Processor sockets per rack with integrated dual plane switching, power, and cooling. The E-cell consists of two 42U high E-racks, which are separated by a cooling rack. There are three compute node types supported: HPE XA730i Gen10 Server is a quad-node Intel Xeon Processor with 2 CPU sockets per node, HPE XA780i Gen10 Server is a single 2 CPU socket node Intel Xeon Processor compute tray with support for up to 4 NVIDIA Tesla for SXM2 GPUs with NVLink and HPE XA760i Server is a quad node Intel Xeon Phi processor compute tray with one CPU socket per node.

Compute Tray Enclosure is 10.5U and provides power, cooling, system control, and network fabric for up to nine compute trays via an integrated midplane. The HPE SGI 8600 System can be expanded by simply adding enclosures, with up to four compute tray enclosures per rack. The enclosure is also designed to support future processor and compute tray technologies.

HPE SGI 8600 System supports both InfiniBand and Intel Omni-Path interconnect technologies with complete flexibility in topology.

The HPE Apollo a6000 Chassis provides power, cooling, and I/O infrastructure to support HPE ProLiant XL Servers. This 5U chassis holds up to 10 hot-swap server trays vertically and fits in a standard rack. The modular HPE Apollo 6000 Chassis can accommodate up to ten server trays, with the flexibility to choose from various trays, HPE ProLiant XL250a Gen9 Server with accelerators, HPE ProLiant XL230a Gen9 Server or a combination of server trays. With newest (summer 2017) Apollo 6000 Gen10 servers performance figures are going to be 323 TFlops/per rack, 1.5TB flash-backed persistent memory and 100GB node-to-node cluster connectivity.

The HPE Apollo 6500 System for deep learning supports up to eight 300W GPU or coprocessors delivering increased performance.

The system consists of three key elements: the HPE ProLiant XL270d Gen9 Server tray, the HPE Apollo d6500 Chassis, and the HPE Apollo 6000 Power Shelf. The HPE ProLiant XL270d Gen9 Accelerator Tray provides up to 168 TFlops of peak half precision performance per server, and up to 37 TFlops of peak double precision performance with eight NVIDIA Tesla P100 and two Intel Xeon E5-2600 v4 processors in a 2U server.

HPe Apollo 4500 are 4U form factor servers that can have up to 2 Intel® Xeon® E5-2600v4 series processors per server, which range from 6-20 cores and can reach up to 135 W and up to 1024 GB DDR4 memory per server.

For object storage, the ultra-dense HPE Apollo 4510 includes one server and up to 68 LFF drives in a 4U chassis for a maximum of 544 TB per system, which equates to over 5.4 PB per 42U rack.

For clustered storage environments, the HPE Apollo 4520 offers two servers with built-in failover capability. The Apollo 4520 offers internal cabling for failover, plus massive disk density of 23 LFF drives per server.

The HPE Apollo 2000 System accommodates up to four independent, hot-pluggable 2P servers in just 2U of rack space. The HPE Apollo 2000 System offers a dense solution with up to four HPE ProLiant server nodes in a standard 2U chassis.

Technically almost the same as Apollo 2000, a new line of products starting with Apollo 70 was introduced to demonstrate the readiness of the ARM platform in HPC from HPE. HPE spent the last year in a pre-production programme with both vendors (ARM, Redhat, Mellanox, SUSE, and others) and users (Argonne, Los Alamos, Sandia labs, and others) to create a complete HPC ARM-based solution. Part of the programme was to develop and optimize different layers from firmware, BIOS, OS up to compilers, libraries, batch schedulers and optimized versions of popular open source HPC software. This platform is not yet 100% comparable to the Intel X86 as it lacks NVIDIA GPU support, but AMD S9510 GPUs are already supported. The CPU is Cavium Thunder-X2 and a server comes in dual socket configuration with 16 DIMMs. Currently, only Infiniband and Ethernet are supported as interconnects.

Apollo 10 series is an entry-level deep learning.

The HPE Apollo sx40 Server is a 1U dual socket server featuring up to four NVIDIA Tesla GPUs (P100) with the high-bandwidth, energy-efficient interconnect NVIDIA NVLink in SXM2 form factor and based on the Intel Xeon Processor Scalable Family (Two Intel Xeon Skylake processors). NVLink enables increased GPU performance for deep learning workloads.

4.5 Lenovo

Lenovo NeXtScale System provides both air-cooled and water-cooled offerings.

The NeXtScale system comprises compute, storage and acceleration nodes in an energy-efficient, low-cost 6-bay enclosure.

Lenovo NeXtScale System M5 Water Cool Technology (WCT) uses direct water-cooling for CPUs, memory and I/O cards. Water is delivered directly to the server and circulated through cooling tubes, supporting water inlet temperatures up to 45 degrees Celsius.

Lenovo nx360 M5 WCT compute tray of 1U is full wide and hosts two half wide server nodes that are cooled by Water Cool Technology (WCT). The platform features Intel Xeon processor E5-2600 v4 series up to 22 cores (enable as much as 528 cores per 6U enclosure). Processors are allowed to run in continuous Turbo mode due to efficient WCT cooling.

The Lenovo NeXtScale n1200 WCT Enclosure utilizes WCT technology to cool six full wide nx360 M5 WCT compute trays for a total of 12 servers per 6U enclosure (up to 84 compute servers

per rack). Designed for water-cooling this enclosure requires no internal fans. It connects to water manifolds that manage inlet and outlet water flows directly to each compute node.

4.6 IBM

The IBM Power System S822LC for High-Performance Computing has Power 8 CPU and 4 NVIDIA Tesla P100 GPU's with NVIDIAs NVLink being used between GPUs. The IBM Power System S822LC provides 2× POWER8 CPU's (2x 8-core 3.25 GHz POWER8 or 2× 10-core 2.86 GHz POWER8). 32 DIMM sockets delivered by 8 memory daughter cards for up to 1TB of memory (512 KB L2 cache per core, 8MB L3 cache per core, up to 64MB per socket and supports 4-32GB DDR4 modules from 128GB to 1TB total memory). NVLink delivers 90 GB/s link speed from GPU to GPU and CPU to GPU. Optional NVMe storage for fast storage input/output. Processor to memory bandwidth 115 GB/s.

The standard backplane includes 2 small form factor (SFF) bays for hard disk drive (HDD) or solid-state disk (SSD). Software RAID. Three PCIe Gen3 slots: two x16 plus one x8 PCIe Gen3, all CAPI enabled.

The IBM PowerAI Deep Learning Frameworks are tuned for use with IBM Power Systems. The fourth release of the PowerAI Deep Learning Frameworks is based on the use of Ubuntu 16.04 on IBM POWER with NVIDIA CUDA 8 and cuDNN v5.1 packages running on HPC hardware.

IBM has announced POWER9 CPUs.

4.7 NEC

4.7.1 NEC Aurora Vector Engine

"Aurora" is the code name for a vector computer that fits on a PCIe card. Vector processor card works like an offload engine for standard x86 Linux server.

SX-Aurora Vector Engine embeds 8cores@1.6GHz, peak performance 2.4 TFlops DP, up to 48 GB HBM2, memory bandwidth 1.2 TB/s.

As a comparison, for HPL, 1 VE is equivalent to a dual socket Intel Skylake node and half a GPU NVIDIA, for STREAM, 1 VE is equivalent to 5 nodes dual socket Skylake and 1,3 Volta.

The workflow can start on an x86 system and, unlike accelerators, the entire application is passed to the vector engine. The x86-system supports the vector engine like a frontend, taking all the workload that does not relate to the application, daemons, and administrative processes. The performance is 300 GFlop/s per core. Interconnect throughput is at PCIe performance, i.e. 16 GB/s.

4.7.2 NEC SX-series: The Next Generation Vector System SX-ACE

Based on Aurora Vector Engine, NEC has announced at SC17 SX-Aurora TUBASA (translation: "Wings"). SX-Aurora TSUBASA is available in five hardware platforms (A100-1 to A100-5), from the tower with 1 VE, to the supercomputer with 64 VE.

The high-end model A100-5 can host up to 64 vector engines (type 10A or 10B), in a proprietary rack. Peak performance is 157 TFlops, with a cumulated bandwidth memory of 76.8 TB/s, max memory capacity is 3 TB.

The system is based upon 8 Vector Hosts (VHs) embedding each 8 VEs managed by 2 Intel Xeon processors (6100 family) with 384 GB memory. VHs are infiniband EDR connected. A100-5 is water and air cooled and HPL power consumption is 30 kW.

4.8 Huawei

Huawei has 20 installations in the current TOP500 [1] list.

Huawei FusionServer X6000 is a high-density server developed for the data centre scale-out architecture with four compute nodes in a 2U chassis with up to twenty-four 2.5-inch NVMe SSDs. It includes modular design and hot-swappable key components such as hard disks, fan modules and power supply units (PSUs), and the operating temperature range is 5-35 °C. A unified management port for managing the entire chassis and all nodes is implemented.

The X6000 server chassis accommodates four XH321 V3 server nodes. These half-width server nodes support Intel Xeon E5-2600 v3/v4 series processors and 16 DDR4 DIMMs. The server node also supports up to six 2.5-inch NVMe SSDs and delivers 4.8 million IOPS.

4.9 Sunway TaihuLight

The Sunway TaihuLight supercomputer, launched in late 2015, is supported by the National High Technology Research and Development Program of China. The supercomputer is developed by the National Research Center of Parallel Computer Engineering & Technology. The Sunway TaihuLight performance numbers are: 125 PFlops of peak performance, 93 PFlops of sustained Linpack performance and 6.05 GFlops/W of Performance-Per-Watt.

Sunway TaihuLight supercomputer’s computing node includes one SW26010 many-core processor, 32 GB memory, node management controller, power supply and interface circuits. Each processor has 4 Management Processing Elements (MPE) and 256 Computing Processing Elements (CPE). The system has 40960 computing nodes of which 256 nodes are integrated into a tightly coupled super-node using fully connected crossing switch to support compute-intensive, communication-intensive and I/O-intensive jobs. Both the MPE and CPE use the SW-64 instruction set and support up to 256-bit vector instructions. The core frequency of the MPEs and CPEs is 1.45GHz and the peak performance of double precision floating point is 3.06 TFlops. The maximum memory bandwidth is 136.51 GB/s.

The network has three levels, central switching network at the top, the super-node network in the middle and the resource-sharing network at the bottom. The central switch network is responsible for the communication between super-nodes. The super-node network is responsible for the all-to-all communication of 256 computing nodes. The resource sharing network is responsible for the all to all communication of 256 Sunway processors in the resource sharing pool. The bisection network bandwidth is 70TB/s with a network diameter of seven.

Sunway TaihuLight cabinets of the computing and network system use indirect water-cooling. The peripheral devices use air and water exchange and the power system uses forced air-cooling. The cabinets use closed-loop and indirect parallel flow water cooling technology.

4.10 Sunway Micro

Equipped with two Chinese SW26010 heterogeneous many-core processors, each processor contains 260 cores divided into four core groups. Each Sunway CPU can support four MPI processes within the corresponding core group. The network-on-chip design boosts the performance of communication between each MPI process. Two processors have high IO and high-speed expansion slot independently. Performance: 6TFlops. Memory: 64-256GB. Storage: 12TB.

5 Data storage and services

The main components of data infrastructure used by HPC systems include:

- scratch storage;
- data management before and after computations;
- preservation/long-term storage for processing input and output data;
- collaboration outside the HPC environment (sync & share).

The data infrastructure importance is growing with the increased data capacity. It becomes increasingly important in both HPC and cloud environments.

5.1 Storage Solutions

With increasing power of CPUs and performance of memory and I/O controllers, the requirements on the storage system are growing as well. This applies to all aspects of storage and data management systems including bandwidth, IOPS and latency. In addition, growing popularity of the data-oriented research, storage and data management systems highlights the importance of the storage systems in the overall HPC software and hardware stack. This results in increased capacity and reliability requirements as well as the need to provide extended functionality such as accessing the data from data analytics environments, ensuring data protection and backup at a relevant, large scale.

The following analysis will cover performance requirements vs storage systems in modern HPC and latest development in this aspect as well as present the broader view on the data management in HPC including big data, data protection and long-term preservation.

5.1.1 Storage performance requirements

'Classical' storage system components in HPC include so-called 'home' space, a persistent, shared storage tier and 'scratch', a high-performance space with parallel access (i.e. ability to share files among job threads). Technologies that implement these spaces must meet several requirements.

Firstly, an HPC system requires a high performance storage system, which is compatible with the performance of the entire computing system. It finally means that the thread performance must be relevant to HPC jobs and workload requirements. This includes high throughput (GB/s), IOPS (number of I/O operations performed in a second) and latency (responsiveness of a storage system to I/O, expressed in milliseconds).

Secondly, HPC jobs are typically highly parallel, so they require support for a large number of simultaneous I/Os. Embarrassingly parallel tasks (e.g. array multiplication, image rendering, genomics pipelines) may produce a high number of sequential and streaming I/O workloads. This I/O traffic must be served efficiently at a range of Gigabytes/s with high parallelism. Massively parallel or tightly coupled workloads may in turn perform numerous parallel I/Os including reads and 'conflicting' writes to the same files, data objects and their areas, which requires locking for data consistency. This puts high stress on the meta-data components of the HPC storage systems.

Thirdly, HPC workloads may produce a mixture of sequential and streaming large block I/O (loading input data to memory and writing/reading the intermediate outputs to/from scratch or saving the final results to persistent storage) and small I/O (in various phases of processing) often combined with locking and file system-level operations, e.g. file or directory creation, truncation. The HPC storage technology must address these mixed I/O specifics of HPC jobs.

Finally, the HPC storage systems must also enable alignment of stripe size to the application I/O size, support customizing striping mode as well as make it possible to efficiently store and access various data structures, e.g. specific file layouts such as HDF5 or NetCDF.

5.1.2 Technologies

The greatest breakthrough in the I/O performance has been recently caused by the NVM (Non-Volatile Memory) technology. NVM-based media are replacing SSDs in most demanding workflows for their improved bandwidth and reduced latency. NVM is used as an extension of the RAM memory (see NVRAM concept discussed earlier) and installed in PCIe-based cards directly in compute/application and I/O nodes of the HPC and storage clusters. NVMe (NVM express) is a data transport protocol, released for the first time in 2012, that aims to replace the current storage protocols including SATA and SAS. While the latter supports only a single 32-command queue (SATA) or a limited number of queues (SAS, SCSI), NVMe can hold 64k of queues with 64k commands each. NVMe technology specifics results in a much larger number of IOPS that can be processed at low latency by storage systems backing HPC clusters at decreased CPU usage per I/O and with lower power consumption. Due to its support for higher parallelism, NVMe technology is more suitable to feed the multi-core CPU with data input and output streams. It addresses multi-threaded, large-scale HPC applications better than SCSI, SATA and SAS [55], [56], [57].

Notably, at the same time, SSDs are replacing SAS and nearline SAS drives due to customer and industry push for high performance and low latency I/O. The capacities of SSD drives grow to several TBs per 2.5” drive form or PCI-based unit (several dozens of TB per unit to appear on the market in 2017/2018) and their manufacturing prices drop due to scale effect: the volume of SSD produced is constantly growing, and more and more vendors are manufacturing NAND flash- based media (as opposed to magnetic drives that are in fact mostly produced by three vendors: HGST, Samsung and Seagate).

Burst buffer technology is a fast and intermediate storage layer positioned between the front-end computing processes and the back-end storage systems. It emerges as a timely storage solution to bridge the ever-increasing performance gap between the processing speed of the compute nodes and the Input/output (I/O) bandwidth of the storage systems. Burst buffer is built from arrays of high-performance storage devices, such as NVRAM and SSD. It typically offers from one to two orders of magnitude higher I/O bandwidth than the back-end storage systems (Figure 28)

An example of the implementation of that technology is the DDN IME. This solution creates a new application-aware fast data tier that resides right between compute and the parallel file system to accelerate I/O, reduce latency and provide greater operational and economic efficiency.

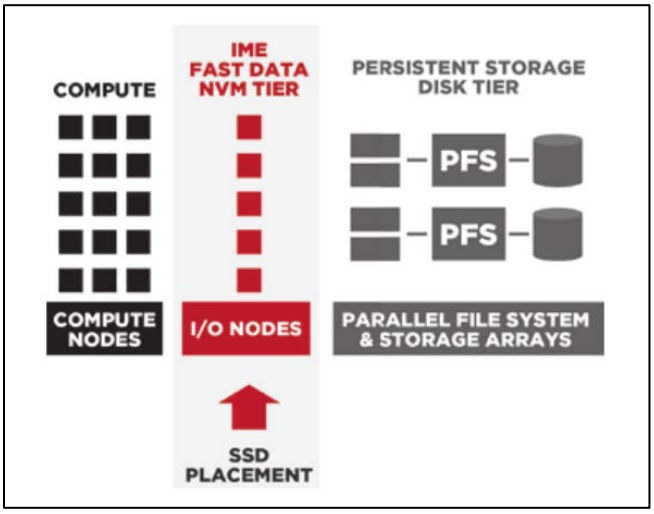

Figure 28. Burst buffer technology

It does this with IME software which has a server and a client component. Rather than issuing I/O to a parallel file system client, the IME client intercepts the I/O fragments and issues these to the IME server layer which manages the NVM media and stores and protects the data. Prior to synchronizing the data to the backing file system, IME coalesces and aligns the I/O optimally for the file system. The read case works in the reverse: file data is ingested into the cache efficiently in parallel across the IME server layer and will satisfy reads from here in fragments according to the read request (Figure 29).

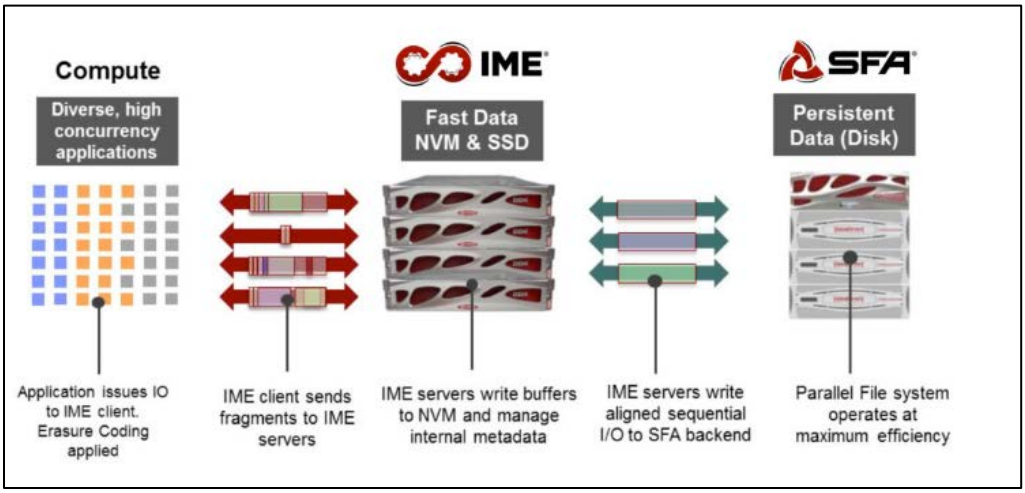

Figure 29. IME technique

The CRAY-Storage (Ex-SEAGATE) implementation for SSDs for intermediate storage acceleration, is based upon an enhanced management of small size I/O:

SEAGATE NYTRO XD:

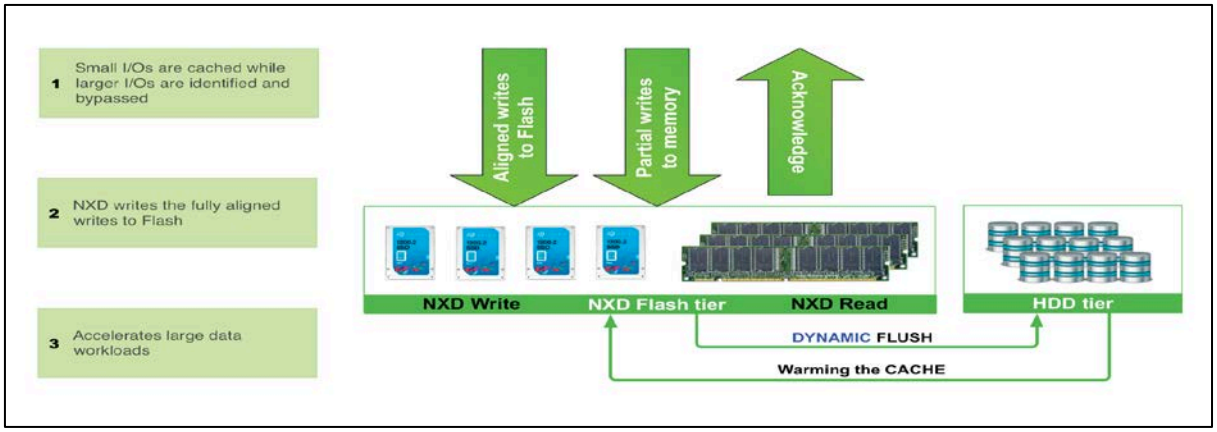

Figure 30. NytroXD small I/O acceleration by CRAY-Storage

Small size blocks are intercepted to be stored on SSD, improving significantly the number of IOPS and random access, without any application impact.

Along with the Intel Xeon Cascade Lake CPU series, new Intel Apache Pass technology has emerged. This is a mechanism for storing data in the server's DIMM (DDR4) memory (Figure 31).

The capacity of the memory module will be 128, 256, or 512 GB. The speed of this solution oscillates at the border of 2666 MT / s. The maximum size of the Apache Pass platform will be up to 6TB (3TB per CPU).

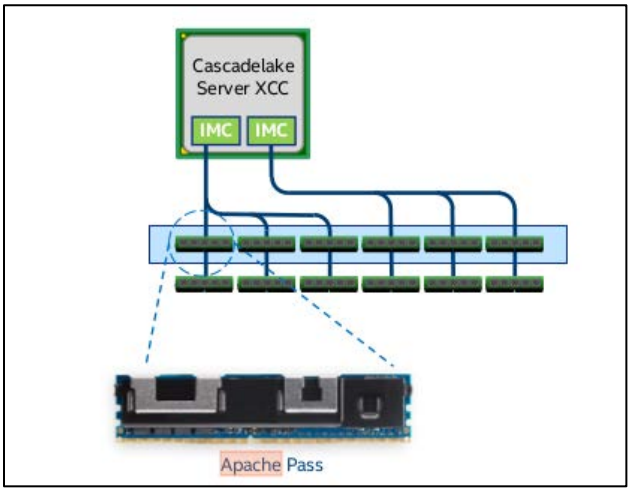

Figure 31. Intel Apache Pass technology

ARM is a family of reduced instruction set computing (RISC) architectures for computer processors, configured for various environments. ARM microservers are implemented according to the systems-on-chips (SoC) and systems-on-modules (SoM) architectures and incorporate memory, interfaces, radios, etc. Processors that have a RISC architecture typically require fewer

transistors than those with a complex instruction set computing (CISC) architecture (such as the x86 processors found in most personal computers), which improves cost, power consumption, and heat dissipation. For supercomputers, which consume large amounts of electricity, ARM could also be a power-efficient solution.

One of the implementations of the ARM based microserver are solutions from the Cynny Space company. They designed the first ARM server engineered for optimal data management. Server has 8.3 x 3.4 cm size and it is the smallest server fully equipped to store data (Figure 32)

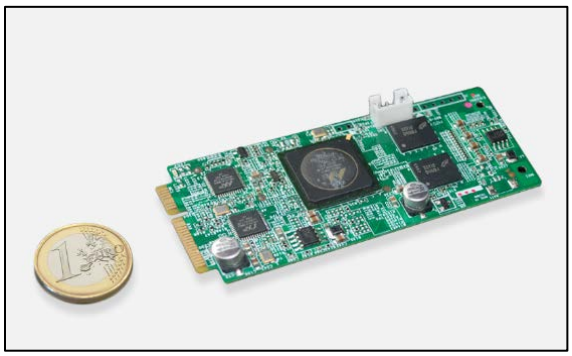

Figure 32. ARM based microserver – Cynny Space

Cynny Space also designed dedicated object storage solutions based on microservers. It is able to offer both high computing power and rack storage density without a single point of failure. The large number of microservers are connected to the network without a layer in between.

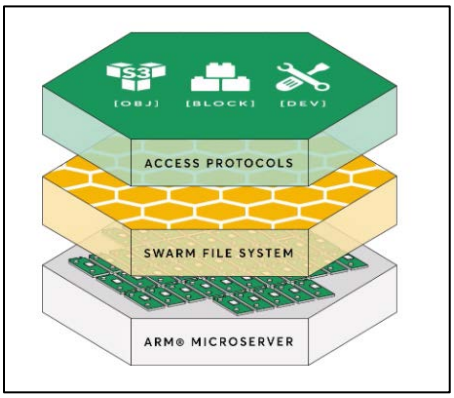

Figure 33. Storage as-a-Service

NVMe over Fabrics (NVMe-oF) is a technology specification designed to enable Non-Volatile Memory express message-based commands to transfer data between a host computer and a target solid-state storage device or system over a network, such as Ethernet, Fibre Channel or InfiniBand.

One of the implementations of NVMe-oF standard is the Mellanox ConnectX cards family. This solution offers a reduction of server CPU utilization (0% in I/O path) and fewer interrupts and context switches (Figure 34).

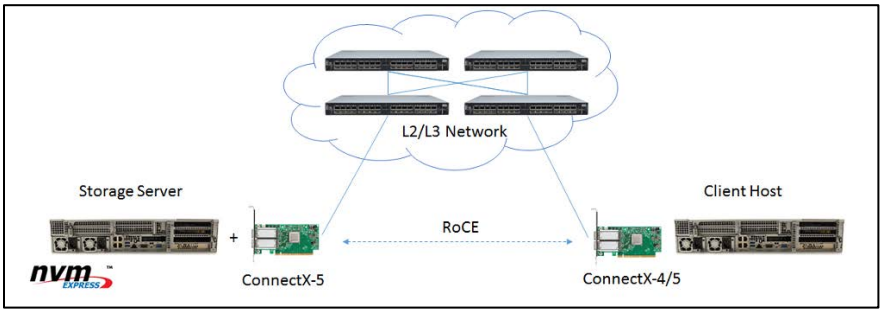

Figure 34. NVM implementation – Mellanox -ConnectX

5.1.3 Storage networking

Important developments have occurred in the area of storage networking. While some technologies are still in their early stage, HPC will be able to benefit from them in the foreseeable future. Among others, NVMe over Fabrics (NVMe-oF), an NVMe interface-based networking technology has been developed [58]. It enables extending communication channels among NVMe-enabled and NVMe storage devices beyond the reach of PCIe interface. Omnipath will also be used in the near future for implementing computing to storage components communication [59], [60].

5.1.4 Shared filesystems

In the area of shared filesystems for HPC, several performance and functionality-related improvements have been made in Lustre. Lustre is the most popular parallel filesystem for HPC available as an open source version, provided by Intel and branded by several vendors including DDN, HP and EMC/Dell (since recently). Major improvements in Lustre include dynamic parallel striping adaptation that extends to static policies supported now - they allow setting per directory or per file striping scheme. Its introduction in planned releases of Lustre will increase possibilities of multi-threaded access to files.

Functionality-wise Lustre is getting extended by enabling NFS / CIFS gateways providing access from those client machines that do not support Lustre clients natively which enables integrating HPC storage with other systems and applications. This provides new possibilities for implementing data management before and after computations, better automation and integration data preservation and long-term storage of computations' input and output data as well as running collaboration outside the close HPC environment (e.g. through sync & share systems such as ownCloud or Seafile).

Providing HDFS gateways, in turn, makes it possible to access data kept in Lustre for data analytics purposes. The latter improvement addresses an important aspect of the recent paradigm shift in data processing. While HPC and HTC keep the positions in several disciplines that require computer-based simulation or solving complex mathematical problems, so-called Big Data uses are becoming relevant. Although data analytics requires a different approach to processing and computation and thus is considered complementary to HPC and HTC, data analytics and HPC and HTC are often performed versus the same datasets or constitute various phases of the overall complex workflows.

Other filesystems are also being developed and improved performance wise and targeted to HPC market. While various Lustre alternatives arise on the market it is important to mention CephFS. The open source CephFS product aims to replace Lustre and overcome its meta-data management limits. While Ceph, especially RADOS (a reliable storage cluster, popular as the back-end for the object storage) and RDB (block storage backing lot of OpenStack deployments) are well established on the marked, important developments happened recently for CephFS towards reaching the production-level maturity and relevant performance (finally keeping the promise of scalable filesystems for HPC). There activities are performed under the coordination of RedHat that keeps the filesystem (and Ceph in general) source code open while providing value-added services, management tools and providing enterprise release of Ceph. Support for RDMA over Infiniband in Ceph is also under development.

Several market players provide parallel filesystems for HPC in a bundle with the disk arrays or disk servers. For instance, DDN offers their SFA boxes integrated with Lustre or GPFS: EXAScaler and GRIDScaler or GS7K. Quanta promotes the QCT QxStor solution composed of RedHat Enterprise Ceph and high-performance and high capacity QuantaPlex and QuantaGrid servers [61], [62], [63].

Another example of integrated software/hardware solution is the EMC Isilon file server based on OneFS filesystem that is becoming prominent in the the HPC market by offering high-performance NFS and CIFS services complemented by HDFS gateways, and integrates enterprise features such as backup and other data management features in the same time. Isilon already supports dual 40Gbit Ethernet interfaces per node (and plans to support 100GbE), so the cluster performance is expressed in bandwidth scales while adding the storage nodes.

While Lustre, CephFS and EMC Isilon/OneFS are mainly addressing large block sequential and streaming I/O, some vendors, including known companies and start-ups, develop and probe the market with IOPS-oriented systems. This includes fully flash arrays, which are not a new concept but they have recently become available with NVMe media in the offering of most vendors. Several interesting technologies, SSD-based, NVMe memory and regular servers are also appearing. An illustrative example of this trend is SANdisk's iON that provides several millions of IOPS at ~50 ms latency, 20+GB/s bandwidth out of the 4U size box including 12-50TB of flash media in PCI cards. Another product (although withdrawn from the market) is the EMC's DSSD solution that is fully NVMe flash array equipped with a specialised interconnect network and PCIe cards to be installed in computing and/or I/O nodes of the HPC cluster. While hardware implementations and software stack of these products differ, a common feature is that they offer a large number of IOPS with a very low latency to/from increasing capacity of SSD or NVMe media.

According to several analysts, the storage hardware market develops towards replacing the magnetic rotating SAS and FC drives with the SSD and (in the more distant future) NVMe media. Several vendors already offer SSD-based storage with pricing similar to SAS based-drives. While this happens for marketing rather than technological reasons, those vendors who do not manufacture rotary drives but produce SSDs and NVMe prefer to sell their own products – the technology migration trend is clearly visible.

The data scratch can be represented by true parallel filesystems. There following are examples:

- Lustre:
  - Enterprise editions available under various brands:
  - Improvements in Lustre: planned dynamic parallel striping adaptation (opposite to static policies supported now - they allow setting per directory or per file striping scheme); this will increase possibilities of multi-threaded access to files
  - NFS / CIFS and HDFS gateways
- EMC Isilon / OneFS – multi-purpose NFS/CIFS/HDFS solution, performance optimised, can address HPC workloads
- Ceph FS
- OrangeFS
- BeeGFS
- Spectrum Scale.

Two additional examples will be described in the following section.

BeeGFS is a free of charge, easy to use, leading parallel and network file system, initially developed at the Fraunhofer Institute for industrial mathematics and formerly known as FhGFS. It allows clients to communicate storage servers via multiple types of interconnects. The more servers added the more performance and capacity are aggregated. BeeGFS is based on the ObjectData and MetaData concept. The former is users' data and the latter is additional information about data, such as access rights, file size as well as the detailed location of the file (e.g. storage server).
BeeGFS consists of the following components (Figure 35):

- ManagementServer (MS) is the component responsible for finding all processes one another. It maintains a list of all file system components – this includes clients, MetaDataServers, MetaDataTargets, StorageServers and StorageTargets,
- ObjectStorageServer (OSS) is the in charge of storing file contents. Each OSS might have one or many ObjectStorageTargets (OST) – where an OST is a RAID - Set (or LUN) with a local filesystem (such as xfs, ext4 or zfs) on top
- MetaDataServer (MDS) stores information about metadata in the system. The architecture of BeeGFS allows a practically unlimited number of MDS. Each MDS has exactly one MetaDataTarget (MDT). The term MDT defines the specific storage device with the appropriate local file system to store MetaData of this MDS. The recommended choice for the file system on MDT is ext4 as it performs well with small files and small file operations.
- the file system client which is a kernel module that has to be installed on all hosts that should access BeeGFS

It is important to note the *server* is used to refer to a Linux process running on a specific machine – not the machine itself.
For production environments, there is also professional support available with defined support response times. The client is published under GPL, and the server is covered by the BeeGFS EULA.

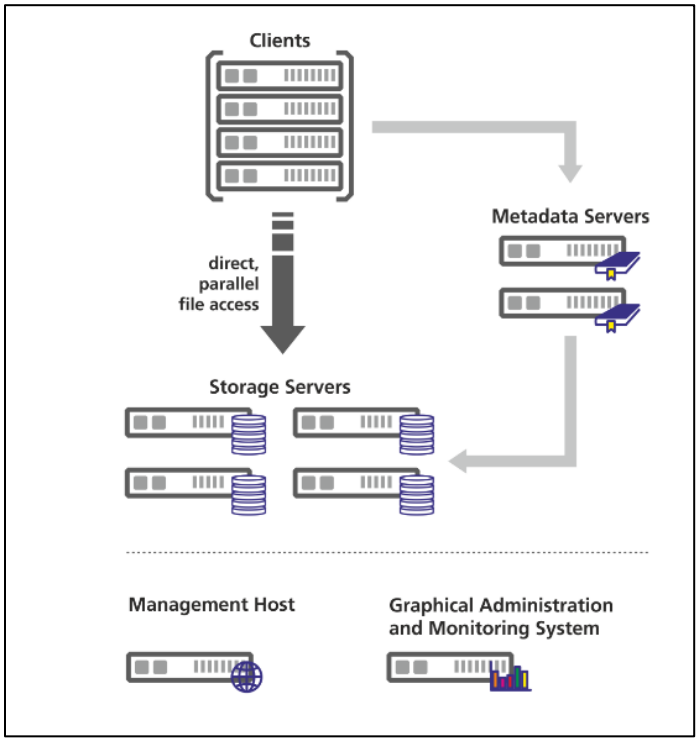

Figure 35. BeeGFS architecture [64]

The Spectrum Scale (SS) from IBM is a scalable, reliable, high-performance and efficient, data and file management solution (based on the IBM General Parallel File System (GPFS) file system, formerly known as Elastic Storage). This enterprise-grade storage management tool takes advantage of the potential of flash storage and automatically transfers data between storage, flash, disk, and tape. The IBM Spectrum Scale reduces storage costs by as much as 90% while contributing to increased security and management efficiency in cloud environments, large data sets, and analysis. The tool also equips users with data-anywhere access that spans storage and locations to accelerate applications across the data centre or around the world. All of this can be managed from a single point using intuitive graphical user interface. With transparent to user storage policies, when applied, data can be compressed or tiered to the tape, flash, disk, cloud or high-performance media. Intelligent caching of data at remote sites ensures that data is available with local read/write performance across geographically distributed sites using data-aware intelligence engine called Active File Management (AFM). IBM Spectrum Scale in the software-defined infrastructure can help improve service, manage risk and rapidly deliver business results at lower costs. The SS file system supports interfaces for file (POSIX, NFS, CIFS), object (S3, SWIFT) or Hadoop Distributed File System (HDFS) for in-place analytics (Figure 36).

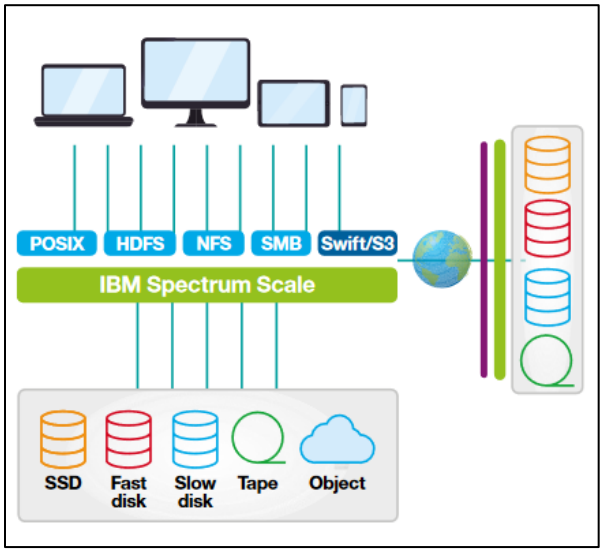

Figure 36. The IBM Spectrum Scale [46]

5.2 Off-line storage

Tape technology has for a long time been seen as a diminishing storage technology, but given the increasing amounts of data being stored and needs for archiving it is not likely that it will be disappearing in the near future. What has been changing is mainly the threshold where tape storage becomes economically feasible. The lowered cost per Tb for disk storage has meant that in order for the low tape media cost to offset the high initial cost for tape libraries you need PB volumes. As a rough rule of thumb 5 PB can be used as a minimum bound for where tape is suitable price wise.

The fact that tapes are off-line is also a factor that is in its favour in some cases. Media can be taken out of the library and stored both off-site and truly off-line to guarantee that it cannot be modified by someone with only electronic access. Write Once Read Many tape media is also available if having unalterable data is more important than being able to reuse tapes.

5.2.1 Tape Drives

Oracle has dropped out of the proprietary enterprise tape market by halting development of the T10k drive technology and concentrating on LTO technology for their libraries. While a large base of existing T10k installations remains, it will slowly become a legacy technology. In effect, this means that IBM is the single source of tape drive technology at the moment, since they are driving the development of both LTO and their proprietary 3592 (aka Jaguar) drives.

The LTO consortium has finalized the LTO-8 standard, and first shipments from vendors were in Q4 2017. Native (uncompressed) capacity has increased to 12 TB from 6 TB with LTO-7. A new capability that will be familiar to users of enterprise tape is that LTO-8 drives can use LTO-7 tapes to store 9 TB, calling the format LTO-8M. This kind of selective formatting of tape media has not been used with LTO drives before. LTO-7 media that has already been used cannot be reformatted, only brand new LTO-7 media can be initialized this way.

IBM introduced the TS1155 drive in mid-2017 with the capability to store 15 TB uncompressed on 3592-JD media. The older TS1150 drives also use JD media, but format them to 10 TB instead. Even older and used JD media can be reformatted by a TS1155 drive in contrast to LTO-8M.

Both LTO-8 and TS1155 drives can transfer uncompressed data at a rate of 360 MB/s, meaning that performance wise the gap between the two technologies is decreasing. Apart from the 15 versus 12 TB data capacity, there are still some other differences that differentiate the two formats. TS1155 drives are available with either Fibre Channel (8 Gbit/s) or Ethernet (10 Gbit/s) connectivity, while LTO-8 drives only use Fibre Channel. Recommended Access Ordering is a feature in 3592 drives to optimize multi-file recalls of data stored on the same volumes. Searching for the right spot on the tape and rewinding is also faster on both 3592 and T10k drives.

Looking at the future, and more specifically the LTO roadmap the aim for coming LTO generations is to double the uncompressed capacity with each generation. Recent LTO generations have been spaced 2-3 years apart, which would yield 24 TB volumes in late 2019 or 2020. IBM has demonstrated prototype tape media capable of storing 330 TB, and while this remains to be commercialized it shows that the potential is certainly there.

5.2.2 Tape Libraries

Competition is much fiercer in the tape library space. For the data amounts handled by HPC sites large automated libraries are needed, and the main competitors are presented in Table 8.

Vendor	Library	Tape Drives	Max Tape Slots
IBM	TS4500	LTO, 3592	23170
Oracle / StorageTek	SL4000	LTO, T10k	9000
	SL8500	LTO, T10k	100880
Spectra Logic	TFinity ExaScale	LTO, 3592, T10k	53460
Quantum	Scalar i6000	LTO	12006

Table 8. List of tape libraries vendors

Several HPC related server vendors are also marketing OEM models of the above, with HPE for example reselling Quantum and Spectra Logic libraries. In some cases, the OEM vendors will reduce the number of available tape technologies.

All library vendors are now supporting background media verification, where tapes in the library are being verified for readability without the applications being aware of this. If this is being done by the library itself or being driven by a server varies among implementations, but all provide a way to get a health status for cartridges in the library. With many thousand cartridges storing data for extended time periods, sometimes without any other read verification, this is becoming a needed feature to ensure data integrity.

5.2.3 LTFS

HPC sites usually have very good network connectivity, but in some cases, the option of ingesting data from external physical media can be useful. Tape can be used for this purpose, and the Linear Tape File System (LTFS) allows the possibility of treating a tape as a kind of portable hard drive. Most commonly it is used for video editing, but also retrieving data from remote measurement sites is an option if network connectivity is lacking. The tape cartridges are sturdy and easy to transport and external tape drives can be attached to workstations or computers attached to instruments.

5.3 Data services

5.3.1 BigData analysis

HPC can be defined as the "Numerical Simulation of Nature" and is used, for example, in Structural and Stability Analysis, Fluid Dynamics, Electrical and Heat Flow, High Energy Physics, Reservoir Simulation, Molecular Modelling and Next Generation Sequencing and many more. Big Data Analysis is about "finding the needle in the haystack", or in a more concrete definition it transforms data into information or does pattern recognition at scale, both in structured and unstructured data and in near real time.

From the HPC view Big Data analysis has an enormous potential to assist HPC in post processing i.e. analysing results of numerical models. For example, by finding correlations in output data resulting from parameter studies or help with tuning and calibrating the numerical model itself and scientists are just starting to develop new use cases that assist their research.

Since right from the beginning of computing there has been a reluctance to delete any data at all, an enormous potential of big data analysis is to turn these "data graves" into useful information

Having said that HPC and Big Data Analysis complement each other very well. However, there is one exception to the rule which comes from the Life Sciences. The analysis phase in Next Generation Sequencing was initially implemented using traditional HPC means. Recently the Broad Institute has published the GTAK (Genomics Analysis Toolkit) implemented in Spark (parallel Hadoop). Thus, in this case, HPC and Big Data Analysis lead to the same results.

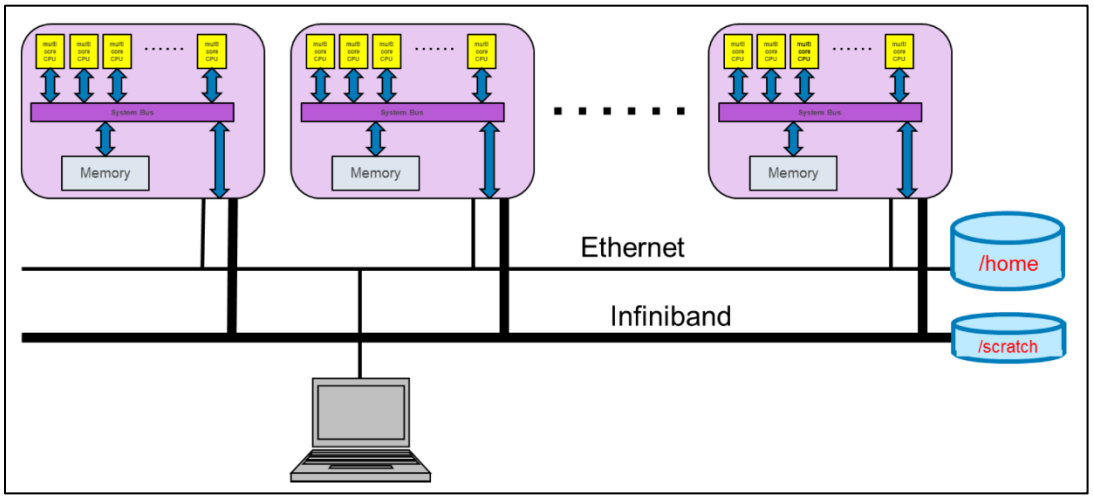

Figure 37. Isilon system offered by Dell EMC (source Dell EMC)

HPC and Big Data Analytics need access to the same data locations. When designing a modern IT environment for both (HPC and Hadoop) this should be a key design criteria. If both environments are designed independently, data access might become the bottleneck in the environment and lead to inefficient execution times. As soon as "Sharing Data means Copying Data" not only a vast amount of time is spent (or wasted) for copy operations of GBs, TBs and sometimes even PBs of data but data gets duplicated which also means extra cost.

As an example presented in Figure 37, the Dell EMC Isilon is addressing exactly the goals of simultaneous access to the same data. Furthermore, Isilion offers NATIVE (optimized and at full performance) access to the same data at the same time using any of the following protocols: NFS, CIFS, HDFS and Swift.

Looking at the compute side, both environments used for intensive computing and analysis of Big Data are built as a cluster of modern multi-CPU / multi-core servers. However, the HPC cluster needs a low latency high bandwidth interconnect (like Infiniband or OPA) while a Hadoop cluster is fine with a lower bandwidth interconnect.

It is essential, however, that both have access to the same data. If built as a combined cluster one could flexibly use nodes either as computing intensive or data analytics, depending on the workload. The cluster could be statically divided or eventually run in a virtual environment. The graphics in Figure 37 shows a possible implementation. From the storage perspective, there is a /scratch high-performance PFS as well as a /home data lake that functions as the central data repository for the HPC and the Hadoop world. The Dell EMC Isilon is an excellent fit for that purpose as shown above.

5.3.2 Machine Learning

Machine Learning (ML) is currently a dynamically developing field of High Performance Computing (HPC) research. It has gained huge successes in a broad area of applications such as speech recognition, computer vision, and natural language processing. With the huge size of data available today, Big Data brings big opportunities and transformative potential for various sectors. On the other hand, it also presents unprecedented challenges to process data and information. As the data keeps getting bigger, machine learning is coming to play a key role in providing big data predictive analytics solutions. In recent years, the models and data available for machine learning applications have grown dramatically. High performance computing offers the opportunity to accelerate performance and deepen the understanding of large data sets through machine learning. Current literature and public implementations focus on small scale environments but small data sets do not scale well in HPC environments due to inefficient data movement and network communication within the compute cluster, originating from the significant disparity in the level of parallelism. Additionally, applying machine learning to extreme scale scientific data is largely unexplored. To leverage HPC for machine learning applications, serious advances will be required in both algorithms and their scalable, parallel implementations.

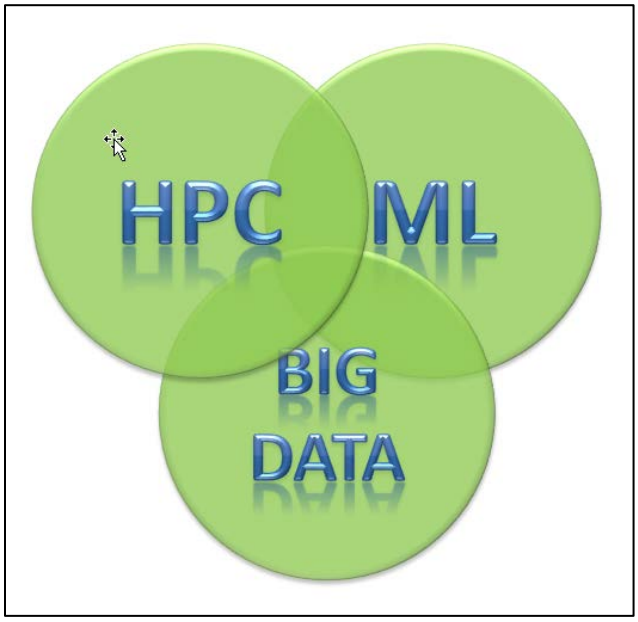

Figure 38. Machine Learning as one of the HPC customer

Machine learning tasks can be divided (according to their characteristics) into the following groups:

- Predicting categorical variables (classification)
- Predicting numerical variables (regression)
- Searching for groupings in the data (unsupervised learning)
- Learning from delayed feedback (reinforcement learning).

6 Paradigm shifts in HPC technologies

6.1 Data Analytics and Artificial intelligence

The application areas “data analytics” and, in particular, “Artificial Intelligence” (AI) are not well defined. In this section, we use these terms to broadly refer to different classes of data-oriented work-loads ranging from discovery of patterns in large-scale data sets to deep learning methods. The interrelation with more traditional HPC is multifaceted:

- Requirements of relevant data analytics and AI work-loads, in particular those involving deep learning, have an increasingly strong impact on technologies, which are also used for or are relevant for HPC.
- The growing requirements for compute resources for some of these work-loads and the need for reducing time-to-solution (e.g., training times) require scale-out approaches and thus adoption of HPC techniques.
- Data analytics and AI work-loads are becoming integral part of scientific computing work-flows.

Data analytics and AI work-loads are furthermore driving the introduction of new or at least improved I/O architectures in order to facilitate fast and typically non-sequential access to vast amounts of data.

6.1.1 Dedicated technologies and architectural features

With the growing importance of the data analytics and AI market, a noticeable impact on hardware developments can be observed. This includes a specialisation depending on the type of work-loads, e.g. optimisation for extremely high throughput of floating-point operations for deep learning applications or optimisation of very high efficiency within a small power envelope for inference applications.

The following trends can be observed:

- Growing support of floating-point arithmetic at reduced precisions (typically FP16) since, for example, typical deep learning methods do not benefit from double precision calculations.
- Introduction of dedicated functional units (e.g. NVIDIA’s Tensor Cores) can be observed that are specifically designed for data analytics and AI work-loads. This trend is enabled by the underlying CMOS technologies, where decreasing transistor feature size (in future 3-dimensional arrangement of transistors) facilitate further increase of the number of transistors per device.
- Market introduction of increasing number of special-purpose devices, which are, e.g., optimised for deep learning or inference tasks, and a growing interest in reconfigurable computing.
- Many of the devices optimised for data analytics and AI can typically not be operated stand-alone but are designed as accelerators. In this context, we observe an optimisation of

processor architectures and work on new or optimised processor bus architectures like PCIe GEN5, OpenCAPI, CCIX and Gen-Z that facilitate integration of such accelerators.

In the following we provide a number of concrete examples for these trends.

In the area of processors, adding support for reduced precision arithmetic has been realised or announced for different architectures. Intel introduced an enhanced version of Xeon Phi under the codename Knights Mill, which mainly differs from the preceding Knights Landing generation by adding support for half-precision floating-point arithmetic. Fujitsu announced that their future ARM-based processor for the Post-K supercomputer will also support instructions for processing vectors with half-precision elements [65]. The other trend for processor products is improved integration of accelerator devices. The most noticeable example is the new IBM POWER9 processor, which features new OpenCAPI interfaces that facilitate high-bandwidth, coherent attachment of accelerators [66].

GPUs are expected to play an important role in further improvement of the computing performance of future supercomputers. Here the impact of AI work-loads is particularly noticeable. Both NVIDIA and AMD pushed the throughput of floating-point operations by improving support of half-precision arithmetic. AMD Radeon Instinct MI25 [67] and NVIDIA V100 [68] devices have a theoretical peak performance of 24.6 and 120 10^{12} FP16 operations/s, respectively. The V100 performance is achieved by additionally introducing so-call Tensor Cores, which perform multiply-add operations on 4x4 matrices.

During recent years a large number of ASICs have been developed with the purpose of accelerating AI work-loads. Google's newest generation of Tensor Processing Units, TPU v2, is optimised for both learning and inference work-loads [69]. New accelerator devices optimised for inference also include Intel Nervana. Fujitsu recently introduced their Deep Learning Unit (DLU) product, which comprises HPC network technology developed for the Post-K computer [70]. Whether these technologies will be integrated into HPC architectures, remains to be seen.

Data analytics and AI workloads are also driving the interest in reconfigurable computing, which in future is expected to play a more important role in HPC. FPGA solutions providers like Xilinx are positioning their products for the data analytics and AI market, mainly by providing IP blocks and software components that are required for integration in established deep learning frameworks like Caffe or TensorFlow [71].

6.1.2 Data analytics and AI in scientific computing workflows

Advances in data analytics and AI facilitate new workflows for scientific discovery. Data obtained from various sources, including experiments and simulations, starts to be processed using advanced data analytics and AI methods to create knowledge, which can be used for improving models used for scalable simulations. An abstract view on such workflows is shown in Figure 39.

Realisation of such workflows could have significant impact on the design and operation of future HPC infrastructures. It requires progress in data management capabilities to facilitate data injection and data access for a broad variety of data sources. Furthermore, a broader variety of infrastructure

services needs to be provisioned to enable workflows to include scalable compute, data analytics and deep learning or inference steps.

Such workflows are being explored in the following areas:

- Materials sciences [72]
- Cancer research and precision medicine [73]

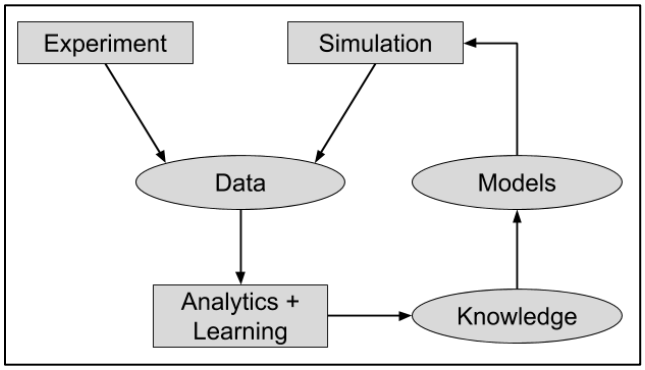

Figure 39: Work-flow combining simulation, data analytics and AI

6.1.3 HPC architectures optimised for data analytics and AI

Various HPC systems, which have recently been deployed or will be available soon, have been designed in a way that they can be used by both scientific computing as well as data analytics and AI applications. The architectures have in common that they use compute accelerators for high-throughput of (possibly lower precision) floating-point operations as well as the integration of non-volatile memory devices to facilitate fast data access. Examples are the following systems:

- JADE: UK Tier-2 resource delivered by Atos and installed at STFC Hartree using NVDIA's DGX-1 as building blocks, each comprising 8 P100 GPUs and SSDs with a capacity of 4 TByte [74].
- Tsubame 3.0: System developed by HP and Tokyo Tech (Japan) comprising 540 nodes, each comprising 4x NVIDIA P100 GPUs and 1x Intel NVMe SSDs with a capacity of 2 TByte [75].
- Summit: System developed by IBM, which is planned to be operational in summer 2018, comprising about 4600 nodes, each comprising 6x NVIDIA V100 GPUs and 1x NVMe SSD with a capacity of 1.6 TByte [76].

The systems listed are based on fat nodes with up to 8 GPUs. With NVIDIA's new NVSwitch technology nodes comprising even larger nodes with up to 16 GPUs are possible [77]. With this technology a scale-up approach becomes possible, but also efforts towards a scale-out of deep learning applications are leveraging HPC technologies. For example, Uber released the Horovod software enabling distributed deep learning in TensorFlow [78].

The US Exascale Compute Project (ECP) has explicitly included deep learning as future exascale challenges and recently released a set of benchmarks targeting cancer research [79].

6.2 Quantum computing

A quantum computer is a computer whose operating physics is governed by the laws of quantum physics. Its main logical unit, called "qubit", is able to physically manifest quantum phenomena, such as superposition, entanglement and tunnelling.

Thanks in particular to the superposition property, which allows each qubit to simultaneously assume the value 0 and the value 1, a quantum computer is able to solve problems of factorial complexity with a single instruction.

By their nature, quantum computers cannot handle all the tasks that are commonly managed by a classic computer (such as I/O and control code). For this reason, their use in the field of HPC turns out to be the most natural choice. In this way, a quantum computer is treated as an accelerator, which can be invoked within a normal classical computer source code.

The interest in quantum computing concerns both large companies in the field of information technology (such as IBM, Microsoft, Google and Intel) and the political world. Some are forming collaborations for the purpose of research: most remarkable are Google and the University of California-Santa Barbara; Lockheed Martin and University of Maryland; and Intel and Delft University of Technology.

The EU recently approved a € 1.13 billion flagship to fund quantum computing research. The Australian government approved in 2016 a five-year loan of 25 million dollars for the development of quantum circuits. In China, they are carrying out interesting research in the field: a few months ago (June 2017) a group of researchers established for the first time an entanglement bond between two photons at 1243 km of distance from each other [80].

There are several companies currently involved in quantum computing. A comprehensive list is available online [81]. At the moment, only the Canadian company D-Wave is available to sell a quantum computer: their latest machine, the D-Wave 2000Q, is available on the market at a price of 15 million dollars.

With regard to the possibility of buying computing hours in the cloud to use quantum machines, there are three companies in particular that are offering this kind of service: IBM, Alibaba and D-Wave itself. The first have recently made universal quantum computers available to users, with 20 and 11 qubits respectively [82, 83].

On 10 November 2017, IBM officially released a quantum "gate-model" (or "universal") computer with 50 qubits. This is an important step forward for quantum computing research: so far, any computation performed on a quantum computer with fewer qubits could be replicated by a sufficiently powerful classical supercomputer. Now it is possible to perform calculations otherwise extremely difficult to perform on classical computers [84].

There are still no articles that show simulations performed with the new quantum computer. Recently, IBM researchers have published an article [85] where they show simulations of small molecules (hydrogen, lithium hydride and beryllium hydride) carried out on IBM quantum computers with a few qubits (up to six).

On the other hand, the other big selling company D-Wave has decided to focus on the quantum annealer, a particular type of quantum computer that does not solve every kind of problem, unlike a universal quantum computer like that produced by IBM, but can solve only a certain class of optimization problems (QUBO problems). This restriction allows them to build machines with a very high number of qubits compared to their competitors. Their last machine on the market implements 2000 qubits: they are already planning the launch of the model with 4000 qubit (expected for 2019).

Another point of strength of the machine built by D-Wave is the ease of use for the programmer: the nature of the machine, i.e. the ability to solve only QUBO problems, allows a simple programming by the user, which consists only in writing the problem of interest in the right form accepted by the machine. Software included in the development kit then allows the conversion of the QUBO problem into machine instructions, directly providing the output to the user. No specific knowledge of quantum mechanics is required to program it (unlike IBM products, for example).

2018 is also an important year for events related to the quantum computing world. From 10th to 12th April the first European user conference "Qubits Europe 2018" sponsored by D-Wave will be held in Munich (Germany). D-Wave has already organized a workshop in Jülich, Germany, held on the 19th and 20th March, with the aim of teaching the use of their new 2000 qubit machine. Also in March (the 21st to be precise), IBM instead organized a conference aimed at publicizing the launch of their new universal quantum computer at 50 qubits, showing technical aspects of the machine and implemented applications already.

6.3 Neuromorphic computing

The term “neuromorphic computing” broadly refers to compute architectures that are inspired by features of brains as found in nature. These features include analogue processing, fire-and-forget communication as well as the extreme high connectivity found in brains of mammals. Neuromorphic computing devices typically belong to the class of non-von Neumann architectures. There is a strong interest in such devices in the context of brain modelling as well as artificial intelligence and machine learning. Benefits of such devices can be speed and energy efficiency.

In this section, we provide an update on the development of and deployment of systems based on neuromorphic computing devices.

6.3.1 BrainScaleS

The BrainScaleS neuromorphic system has been developed at the University of Heidelberg [86]. It uses analogue circuits to implement models of neuronal processes. A special feature is its wafer-scale integration, which allows for very fast communication between the neurons within a wafer. Both features allow for simulations to run several orders of magnitude faster compared to biological speeds. A BrainScaleS wafer consists of 48 reticles, each of which hold eight High-Count Analogue Neural Network (HiCANN) dies. Each of these dies implements 512 neurons and over 100 000 synapses. There are two levels of communication: one within the wafer and one between the wafers. The latter is realised through serial links implemented in FPGAs.

Recently, the BrainScaleS-2 chip was announced that has been developed with support from the Human Brain Project [87]. The new chip features programmable on-chip learning capabilities and a new concept called dendritic computation developed in close collaboration with neuroscientists. A system based on the new chip will become part of the Neuromorphic Computing Platform of the Human Brain Project [88]. The system has so far been mainly used for brain research. In future it is planned to focus more on learning applications.

6.3.2 SpiNNacker

SpiNNaker (Spiking Neural Network Architecture) is an architecture based on simple ARM9 cores mainly developed at Manchester University [89]. The processor is based on a system-on-a-chip design integrating 18 cores as well as the network logic including 6 communication links. The architecture is optimised for brain simulations and for this reason was able to discard usually applied design principles as memory coherence, synchronicity and determinism. Point-to-point communication happens through unreliable fire-and-forget transmissions of small packets.

Recently the project presented a prototype of the new SpiNNaker-2 chip [87]. It features 144 ARM Cortex M4F cores on a single chip.

Large deployments based on SpiNNaker-2 are planned within the Neuromorphic Computing Platform of the Human Brain Project [88]. The architecture so far has been used for modelling of neuronal networks for brain research as well as for robotics interaction.

6.3.3 Loihi

Intel announced in autumn 2017 a neuromorphic chip for learning called Loihi [90]. The chip comprises 128 neuromorphic cores plus 3 standard x86 cores. The neuromorphic cores are interconnected via an asynchronous mesh network that supports a wide range of sparse, hierarchical and recurrent neural network topologies with each neuron capable of communicating with thousands of other neurons. The architecture in total implements about 130,000 neurons and 130 million synapses. The chip was implemented using a very advanced 14 nm process technology.

The architecture is foreseen to be used for learning applications.

6.3.4 TrueNorth

TrueNorth is a reconfigurable processor developed by IBM comprising of 1 million artificial neurons and 256 million artificial synapses, which are organised in 4096 neurosynaptic cores [91]. Like SpiNNacker, the design is digital but asynchronous (except for a clock running at an extremely low frequency of 1 kHz). The chips can be connected directly together to form larger systems. A scale-up configuration has been realised integrating 16 chips on a single board. For an alternative scale-out configuration single-chip boards are interconnected via a 1-gigabit Ethernet network. The design is heavily optimised for power efficiency. A single TrueNorth chip consumes only 70 mW.

A full ecosystem around the TrueNorth hardware is growing and currently in use at more than 30 universities, government agencies and labs. In June 2017 IBM announced that they will deploy a system at U.S. Air Force Research Laboratory that is “equal to 64 million neurons” [92]. The

technology has been applied to a growing number of different areas including signal processing (e.g. video tracking, supernova detection), robotics, neural circuit modelling, and optimisation.

6.4 Heterogeneous systems

In the race towards the Exascale, power consumption is becoming the most prominent factor of performance. If the target for an exaflop machine is to have the power consumption in the range of 20 to 40 MW, the TOP500 list (see Table 9) shows that, using 2017 technology this goal is not achievable.

Rank	Name	Accelerator	Year	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (MW)	Extrapolation to Exa (MW)	Yield in % (Max/Peak)
1	Sunway TaihuLight	Sunway SW26010	2016	93014.60	125435.90	15.371	122.54	74.15
2	Tianhe-2	Xeon Phi 31S1P	2013	33862.70	54902.40	17.808	324.36	61.68
3	Piz Daint	NVIDIA P100	2017	19590.00	25366.30	2.272	89.57	77.23
4	Gouyou	PEZY-SC2	2017	19135.80	28192.00	1.350	47.89	67.88
5	Titan	NVIDIA K20X	2012	17590.00	27112.55	8.209	302.77	64.88
6	Sequoia		2011	17173.22	20132.66	7.890	391.90	85.30
7	Trinity	Xeon Phi 7250	2017	14137.30	43902.57	3.844	87.55	32.20
8	Cori	Xeon Phi 7250	2016	14014.70	27880.65	3.939	141.28	50.27
9	Oakforest	Xeon Phi 7250	2016	13554.60	24913.46	2.719	109.13	54.41
10	K Computer		2011	10510.00	11280.38	12.660	1122.29	93.17

Table 9. November 2017 TOP500 list

This table tells us also that 8 out of 10 of the most powerful machines are accelerated. It is also interesting to notice that the two non-accelerated machines are old machines (2011). Notice also that technology of 2011 was less efficient, in terms of electrical power than the most recent ones, as can be seen through the extrapolation column, yet the Linpack yield of those two machines was higher.

The analysis of the Green500 list [2], shown in Table 10, tells us that the most power efficient machines worldwide are all accelerated utilizing the corresponding hardware.

Green500 rank	TOP500 rank	Name	Accelerator	RMAX (Tflop/s)	Power (kW)	Efficiency (GFlops/W)	RPeak (TFlops/s)	Yield in % (Max/Peak)
1	259	Shoubou system B	PEZY-SC2	842.0	50	17.009	1127.68	74.67
2	307	Suiren2	PEZY-SC2	788.2	47	16.759	1082.573	72.81
3	276	Sakura	PEZY-SC2	824.7	50	16.657	1127.680	73.13
4	149	DGX SaturnV	NVIDIA Tesla V100	1070.0	97	15.113	1819.752	58.80
5	4	Gyoukou	PEZY-SC2	19135.8	1350	14.173	28192.000	67.88
6	13	Tsubame3.0	NVIDIA Tesla P100	8125.0	792	13.704	12127.069	67.00
7	195	AIST AI Cloud	NVIDIA Tesla P100	961.0	76	12.681	2148.800	44.72
8	419	RAIDEN GPU	NVIDIA Tesla P100	635.1	60	10.603	947.712	67.01
9	115	Wilkes-2	NVIDIA Tesla P100	1193.0	114	10.428	1751.616	68.11
10	3	Piz Daint	NVIDIA Tesla P100	19590.0	2272	10.398	25326.264	77.35

Table 10. November 2017 Green 500 list

Table 11 shows that only a fraction of the peak performance is used in practice, as soon as the application deviates from a “Linpack” like (i.e. dense linear algebra) type of workload. As such, the HPCG benchmark emphasizes the importance of the memory bandwidth and therefore the balance of the system.

Rank	Name	Rmax (TFlop/s)	Rpeak (TFlop/s)	HPCG (TFlop/s)	Yield in % (HPCG/Rmax)
1	Sunway TaihuLight	93014.594	125435.904	480.800	0.52
2	Tianhe-2	33862.700	54902.400	580.109	1.71
3	Piz Daint	19590.000	25326.264	486.398	2.48
4	Gouyou	19135.800	28192.000	N/A	N/A
5	Titan	17590.000	27112.550	322.322	1.83
6	Sequoia	17173.224	20132.659	330.373	1.92
7	Trinity	14137.300	43902.566	546.124	3.86
8	Cori	14014.700	27880.653	355.442	2.54
9	Oakforest	13554.600	24913.459	385.479	2.84
10	K Computer	10510.000	11280.384	602.736	5.73

Table 11. Effective yield of the Top10 machines

From those tables we can draw three important conclusions:

1. To build even more powerful computers at a reasonable electrical power, the architects will have to introduce some kind of heterogeneity in their design.
2. The yield (achievable versus peak performance) will likely diminish over time prompting to innovative solutions to overcome this performance loss.
3. The balance of the final design is important to get a general-purpose computer that allows the use of a decent fraction of the peak performance. Furthermore the solution must be usable i.e. programmable at a decent cost.

The heterogeneity is (and will be) observed at different levels:

- Rather having a single homogeneous computer node, the machine itself can be split in two (or more) subparts that are devoted to different classes of workloads. This is, for example, the choices made for Trinity at LANL [93] or Tera 1000 at CEA [94], where a smaller part of the machine is Haswell based and the bulk of the compute power comes from a larger part based on Xeon Phi 7250 (Knights Landing [95]). We predict that this will be the preferred architecture choice in the future.
- The memory hierarchy can be augmented by using stacked memory. Latest generations of GPUs are using HBM2 stacks. The KNL is relying on its MCDRAM (HMC from Micron) to boost the memory accesses along with its DDR4 that provides the bulk of the storage. Our (educated) guess is that we will see most HPC oriented CPU designs relying on HBM2 (or higher) in the near future to overcome the DDR bandwidth limitations. The question that has no answer, as of today, is whether the DDR will still be present along with the HBM (which will serve as a, possibly manual, cache) or will simply disappear leading to simpler architecture for processor designers. This question can be extended to the usage of non-volatile memory (NVM) in conjunction or replacement to DDR. Having NVM + DDR + HBM is another form of heterogeneity.
- At the node level, CPUs will coexist with some sort of accelerators. The nature of the accelerator will depend on the type of workloads. Currently, the most popular solution is the GPU. The Green500 demonstrates that other less classical options are possible. One active track of investigation is to see what can be done with FPGAs (see the EuroExa project [96]). Neural network processors are gaining momentum [97] and will certainly take place in supercomputers to speed up some compute phases. Further along the road, quantum chips will also take a place in our computers: the current implementation of the Shor's algorithm runs on a classical CPU and accelerates parts of its computations on a quantum device [98]. It won't be surprising to see machines with different flavours of those accelerators in a not so distant future.