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

D5.2

Market and Technology Watch Report Year 2
Final

Version: 1.01
Author(s): Aris SOTIROPOULOS (GRNET)
Date: 21.03.2019

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: D5.2</td></tr>
<tr><td colspan="2">Deliverable Nature: Report</td></tr>
<tr><td rowspan="2">Dissemination Level:
PU*</td><td>Contractual Date of Delivery:
31 / March / 2019</td></tr>
<tr><td>Actual Date of Delivery:
29 / March / 2019</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 2	
	ID:	D5.2	
	Version:	1.01	**Status:** *Final*
	Available at:	http://www.prace-ri.eu	
	Software Tool:	Microsoft Word (Windows and Mac)	
	File(s):	D5.2.docx	
Authorship	**Written by:**		Aris SOTIROPOULOS (GRNET)
	Contributors:		Adem Tekin (UHEM) Ahmet Tuncer Durak (UHEM) Andreas Johansson (LiU) Aris Sotiropoulos (GRNET) Dirk Pleiter (FZJ) Eric Boyer (CINES) Filip Stanek (IT4I) François Robin (CEA) Gert Svensson (KTH) Guillaume Colin de Verdière (CEA) Jean-Philippe Nomine (CEA) K.Wadówka (PSNC) Kalle Happonen (CSC) Marcel Bruckner (BSC) Mateusz Tykierko (WCSS) Michael Stephan (FZJ) Mirko Cestari (CINECA) Norbert Meyer (PSNC) Panayiotis Tsanakas (GRNET) Pawel Woszuk (PSNC) Philippe Segers (GENCI) Samuli Saarinen (CSC) Stephane Requena (GENCI) Susanna Salminen (CSC)
	Reviewed by:		Walter Lioen (SURFsara) Thomas Eickermann (FZJ)
	Approved by:		MB/TB

Document Status Sheet

Version	Date	Status	Comments
0.01	09/January/2019	Draft	Aris Sotiropoulos (GRNET) – Initial version, TOC, Chapter Leaders, Contributors
0.02	10/January/2019	Draft	Adem Tekin (UHEM), Ahmet Tuncer Durak (UHEM) – Chapter 2.1
0.03	15/January/2019	Draft	Guillaume Colin de Verdière (CEA) – Chapters 7.1 and 7.2
0.04	16/January/2019	Draft	Mirko Cestari (CINECA) – Ch 3.9 PPI4HPC
0.05	17/January/2019	Draft	Aris Sotiropoulos (GRNET), Panayiotis Tsanakas (GRNET) – Ch. 2.3 Business analysis
0.06	18/January/2019	Draft	Andreas Johansson (LiU) – Ch 5.2 Off-line storage
0.07	20/January/2019	Draft	Mateusz Tykierko (WCSS) – Ch 2.5 Consolidation in HPC market
0.08	21/January/2019	Draft	Dirk Pleiter (FZJ) – Ch 4.1 Processors, Ch 6.2 IBM, Fujitsu
0.09	21/January/2019	Draft	Susanna Salminen (CSC) – Ch 6.3 Cray
0.10	22/January/2019	Draft	Eric Boyer (CINES) – Chapter 2.2 Exascale Initiatives
0.11	22/January/2019	Draft	Dirk Pleiter (FZJ) – Ch 7.3 Neuromorphic computing
0.12	22/January/2019	Draft	Jean-Philippe NOMINE (CEA), Mirko CESTARI (CINECA), François ROBIN (CEA), Dirk PLEITER (FZJ), Philippe SEGERS (GENCI) – Ch 3 EU HPC Landscape and Technology Update
0.13	23/January/2019	Draft	Marcel Bruckner (BSC) – Ch 4.2.1 GPU
0.14	24/January/2019	Draft	Michael Stephan (FZJ) – Ch 6.6 Lenovo
0.15	24/January/2019	Draft	Mateusz Tykierko (WCSS) – References for Ch 2.5
0.16	24/January/2019	Draft	Marcel Bruckner (BSC) – Ch 4.4 Interconnect
0.17	25/January/2019	Draft	Filip Stanek (IT4I) – Ch. 4.3 Memory Technologies, Ch. 4.4 Interconnects
0.18	25/January/2019	Draft	Andreas Johansson (LiU) – Ch 5.1 Storage Solutions
0.19	28/January/2019	Draft	Mirko Cestari (CINECA) – Ch 4.1.1 x86_64 processors
0.20	29/January/2019	Draft	Andreas Johansson (LiU) – Ch 5.1 Storage Solutions++
0.21	13/February/2019	Draft	François Robin (CEA) – Various changes
0.22	17/February/2019	Draft	Filip Stanek (IT4I) – Ch. 4.4 Interconnects, Ch. 6.1 HPE
0.23	18/February/2019	Draft	Mirko Cestari (CINECA) – Ch 1 Introduction
0.24	18/February/2019	Draft	Samuli Saarinen (CSC) – Ch 6.5 Atos
0.25	18/February/2019	Draft	Norbert Meyer, K.Wadówka, Pawel Woszuk (PSNC), Marcel Bruckner (BSC) – Ch 5 Data Storage and Data Management
0.26	19/February/2019	Draft	Samuli Saarinen (CSC) – Ch 2.4 Cloud computing in HPC
0.27	27/February/2019	Draft	Aris Sotiropoulos (GRNET) – Ch 6.4 NEC, Acronyms and Abbreviations
0.28	28/February/2019	Intra-WP5 review	François Robin (CEA) – Executive Summary and Conclusion
0.29	28/February/2019	WP5 rev	François Robin (CEA) – Corrections/comments
0.30	1/March/2019	WP5 rev	Philippe Segers (GENCI) – Contribution to Executive Summary and Conclusion, populating keywords section
0.31	1/March/2019	WP5 rev	Gert Svensson (KTH) – Multiple corrections
0.32	4/March/2019	WP5 rev	Dirk Pleiter (FZJ) – Ch 4.1.2 ARM processors additions/corrections
0.33	5/March/2019	WP5 rev	Adem Tekin (UHEM), Ahmet Tuncer Durak (UHEM) – Ch 2.1 corrections
0.34	5/March/2019	WP5 rev	Gert Svensson (KTH) – Multiple corrections, Susanna Salminen (CSC) – Corrections
0.35	6/March/2019	WP5 rev	Gert Svensson (KTH) – Corrections, Dirk Pleiter (FZJ) – Corrections
0.36	6/March/2019	WP5 rev	Eric Boyer (CINES) – Chapter 2.2 review and add-ons
0.37	7/March/2019	PRACE r	Aris Sotiropoulos (GRNET) – prepare for PRACE internal review

0.40	15/March/2019	PRACE r	Aris Sotiropoulos (GRNET) – initial corrections after comments from PRACE reviewers (Walter Lioen (SURFsara), Thomas Eickermann (FZJ))
0.41	19/March/2019	PRACE r	Mateusz Tykierko (WCSS), Filip Stanek (IT4I), Susanna Salminen (CSC), Dirk Pleiter (FZJ), Mirko Cestari (CINECA), Adem Tekin (UHEM), Eric Boyer (CINES), Stephane Requena (GENCI) – addressing PRACE reviewers' comments
0.42	21/March/2019	PRACE r	Marcel Bruckner (BSC), Kalle Happonen (CSC) – addressing PRACE reviewers' comments
0.43	21/March/2019	PRACE r	François Robin (CEA) – important amendments by WP Leader
1.00	21/March/2019	Final	Aris Sotiropoulos (GRNET) – finalize for TB/MB approval

Document Keywords

Keywords:	HPC, Top500, Exascale, Strategy, Market watch, Technology watch Business analysis, Cloud computing in HPC, EuroHPC, PRACE, ETP4HPC, BDVA, FETHPC, EPI, PPI4HPC, HBP, FENIX, ICEI, Intel, AMD, ARM, POWER processors, FPGA, GPU, NVIDIA, Memory technologies, Interconnect, Data Storage, Data Management, IBM, Spectra Logic, Oracle StorageTek, Quantum tape, Data services, DDN, Seagate, HPE, Fujitsu, Huawei, Cray, NEC, ATOS-BULL, Lenovo, AI, Heterogeneous technologies, Neuromorphic computing

Table of Contents

List of Figures

List of Tables

References and Applicable Documents

[1] https://www.top500.org/

[2] https://www.top500.org/green500/

[3] http://www.hpcg-benchmark.org/

[4] https://www.vi4io.org/io500/start

[5] PRACE-5IP D5.1 Deliverable "Market and Technology Watch Report Year 1", 2018

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

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

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

[9] https://www.olcf.ornl.gov/olcf-resources/compute-systems/summit/

[10] https://computation.llnl.gov/computers/sierra

[11] https://www.nitrd.gov/nsci/

[12] https://www.exascaleproject.org/

[13] https://eurohpc-ju.europa.eu/

[14] https://www.intersect360.com/

[15] https://hyperionresearch.com/

[16] https://kubernetes.io/

[17] https://www.sylabs.io/singularity/

[18] https://github.com/NERSC/shifter

[19] https://www.slideshare.net/AmazonWebServices/the-nitro-project-nextgeneration-ec2-infrastructure-aws-online-tech-talks

[20] https://www.prnewswire.com/news-releases/marvell-technology-completes-acquisition-of-cavium-300676912.html

[21] https://www.broadcom.com/company/news/financial-releases/2357930

[22] https://www.arista.com/en/company/news/press-release/6070-pr-20180912

[23] https://news.hpe.com/hpe-to-acquire-plexxi/

[24] https://news.arubanetworks.com/press-release/hpe-acquire-cape-networks

[25] https://www.anandtech.com/show/12989/mips-acquired-by-wave-computing

[26] https://www.onestopsystems.com/article/oss-acquires-bressner-technology-expands-european-presence

[27] https://www.hpcwire.com/off-the-wire/ibm-to-acquire-red-hat/

[28] https://www.hpcwire.com/off-the-wire/ddn-completes-60-million-tintri-acquisition-and-enters-enterprise-virtualization-market/

[29] https://www.hpcwire.com/off-the-wire/xilinx-announces-the-acquisition-of-deephi-tech/

[30] https://www.hpcwire.com/off-the-wire/storage-leader-ddn-acquires-lustre-file-system-capability-from-intel/ https://www.hpcwire.com/off-the-wire/storage-leader-ddn-acquires-lustre-file-system-capability-from-intel/

[31] https://www.oracle.com/corporate/acquisitions/talari/

[32] https://www.oracle.com/corporate/acquisitions/datafox/

[33] https://www.oracle.com/corporate/pressrelease/oracle-buys-datascience-051618.html

[34] https://news.microsoft.com/2018/06/04/microsoft-to-acquire-github-for-7-5-billion/

[35] PRACE 5IP Deliverable D2.2 "Report on stakeholder management"

[36] https://ec.europa.eu/programmes/horizon2020/en/news/21-new-h2020-high-performance-computing-projects

[37] https://www.etp4hpc.eu/cppp-monitoring.html

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

[39] Council Regulation (EU) 2018/1488 of 28 September 2018 establishing the European High Performance Computing Joint Undertaking ST/10594/2018/INIT

[40] https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32018R1488&from=EN

[41] http://www.prace-ri.eu/postitionpaper-eurohpc-prace-2019/

[42] http://www.etp4hpc.eu/

[43] http://www.etp4hpc.eu/membership.html

[44] http://www.bdva.eu/

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

[46] http://www.bdva.eu/PPP

[47] https://exdci.eu/

[48] https://exdci.eu/about-exdci

[49] https://www.eurolab4hpc.eu/

[50] www.focus-coe.eu/

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

[52] https://exdci.eu/collaboration/coe

[53] http://www.deep-project.eu/deep-project

[54] http://montblanc-project.eu/

[55] https://euroexa.eu/

[56] https://ec.europa.eu/digital-single-market/en/news/european-processor-initiative-consortium-develop-europes-microprocessors-future-supercomputers

[57] https://www.ppi4hpc.eu/

[58] https://ec.europa.eu/digital-single-market/en/public-procurement-innovative-solutions

[59] https://ted.europa.eu/TED/notice/udl?uri=TED:NOTICE:202138-2018:TEXT:EN:HTML

[60] https://www.humanbrainproject.eu/en/

[61] https://fenix-ri.eu/

[62] AnandTech, “Intel shows Xeon scalable gold 6138p with integrated FPGA shipping to vendors” (https://www.anandtech.com/show/12773/intel-shows-xeon-scalable-gold-6138p-with-integrated-fpga-shipping-to-vendors)

[63] HPCwire, “Requiem for a Phi: Knights Landing Discontinued”, (https://www.hpcwire.com/2018/07/25/end-of-the-road-for-knights-landing-phi/)

[64] ALCF, “ALCF Aurora 2021 Early Science Program: Data and Learning Call For Proposals”, (https://www.alcf.anl.gov/alcf-aurora-2021-early-science-program-data-and-learning-call-proposals)

[65] Wikichip, “Intel Cascade Lake Brings Hardware Mitigations, AI Acceleration, SCM Support“ https://fuse.wikichip.org/news/1773/intel-cascade-lake-brings-hardware-mitigations-ai-acceleration-scm-support/

[66] R. Hazra “Intel Keynote at SC2018”

[67] Wikichip, “Cascade Lake AP - Cores - Intel” (https://en.wikichip.org/wiki/intel/cores/cascade_lake_ap)

[68] STH, “Intel Cooper Lake Xeon For Training Details from Architecture Day 2018” (https://www.servethehome.com/intel-cooper-lake-xeon-for-training-details-from-architecture-day-2018/)

[69] AnandTech, “Intel Server Roadmap: 14nm Cooper Lake in 2019, 10nm Ice Lake in 2020“, (https://www.anandtech.com/show/13194/intel-shows-xeon-2018-2019-roadmap-cooper-lakesp-and-ice-lakesp-confirmed)

[70] wccftech, “AMD Zen 2 Based 7nm Rome Server Processors Were Designed To Compete Favorably With Intel Ice Lake-SP Xeon CPUs, Aiming For Multi-Digit Server Market Share Gains” (https://wccftech.com/amd-epyc-rome-zen-2-7nm-cpus-compete-favorably-against-intel-ice-lake-xeons/)

[71] AnadTech, “AMD’s future in servers: New 7000 series CPUs launched and EPYC analysis” (https://www.anandtech.com/show/11551/amds-future-in-servers-new-7000-series-cpus-launched-and-epyc-analysis)

[72] top500, “AMD Notches EPYC Supercomputer Win with Next-Generation Zen Processor” (https://www.top500.org/news/amd-notches-epyc-supercomputer-win-with-next-generation-zen-processor/)

[73] Feldmann M. Feldmann, “China Reveals Third Exascale Prototype”, top500.org, 22 October 2018.

[74] Ampere Computing, “Ampere 64-bit ARM Processor”, product brief (https://amperecomputing.com/wp-content/uploads/2018/02/ampere-product-brief.pdf).

[75] Cavium, “ThunderX2 CN99XX Product Brief”, product brief (https://www.marvell.com/documents/cmvd78bk8mesogdusz6t/).

[76] Huawei, “Huawei Unveils Industry's Highest-Performance ARM-based CPU”, press release (https://www.huawei.com/en/press-events/news/2019/1/huawei-unveils-highest-performance-arm-based-cpu).

[77] Wikichip, “Kunpeng 920 (Hi1620) – HiSilicon”, https://en.wikichip.org/wiki/hisilicon/hi16xx/hi1620.

[78] Toshio Yoshida, “Fujitsu High Performance CPU for the Post-K Computer”, presentation at Hot Chips 2018.

[79] DDN, “DDN Unveils Professional Support for Lustre Clients on ARM-based Platforms”, press release, 12 November 2018.

[80] IBM, “POWER9 Processor User’s Manual”, April 2018.

[81] IBM, “IBM Power System AC922: Introduction and Technical Overview”, IBM Redbook, March 2018.

[82] June 11, 2018, https://www.amd.com/en/press-releases/2018-11-06-next-era-artificial-intelligence-and-high-performance-computing-hpc

[83] November 8, 2018, https://www.top500.org/news/amd-introduces-first-7nm-datacenter-gpus/

[84] September 12, 2018, https://nvidianews.nvidia.com/news/new-nvidia-data-center-inference-platform-to-fuel-next-wave-of-ai-powered-services

[85] September 13, 2018, https://www.top500.org/news/nvidia-unveils-new-inferencing-gpu-for-datacenter/

[86] January 16, 2019, https://insidehpc.com/2019/01/nvidia-t4-gpus-come-to-google-cloud-for-high-speed-machine-learning/

[87] November 7, 2018, https://www.nextplatform.com/2018/11/07/competition-finally-comes-to-datacenter-gpu-compute/

[88] M. Feldman “CPU wars and exascale clarity: HPC in 2019” https://www.nextplatform.com/2019/01/11/cpu-wars-and-exascale-clarity-hpc-in-2019/

[89] Brett Williams, Micron, Meeting at SC18, Dallas, TX, US, 2018

[90] SK Hynix, https://www.techquila.co.in/sk-hynix-develops-first-16-gb-ddr5-5200-memory-chip/

[91] JEDEC, https://www.jedec.org/news/pressreleases/jedec-updates-groundbreaking-high-bandwidth-memory-hbm-standard-0

[92] Micron, http://investors.micron.com/static-files/36fd1838-b055-4e8c-97a4-972d4a1ad0cf

[93] Micron, https://www.micron.com/-/media/client/global/documents/products/product-flyer/nvdimm_flyer.pdf?la=en

[94] July 19, 2018, https://www.hpcwire.com/2018/07/19/infiniband-still-tops-supercomputing/

[95] November 15, 2018, http://www.mellanox.com/solutions/hpc/top500.php

[96] November 12, 2018, https://www.businesswire.com/news/home/20181112005379/en/Mellanox-InfiniBand-Ethernet-Solutions-Accelerate-Majority-TOP500

[97] October 24, 2018, https://insidehpc.com/2018/10/interconnect-future-paving-road-exascale/

[98] Mellanox, http://www.mellanox.com/related-docs/whitepapers/WP_Mellanox_Socket_Direct.pdf

[99] Top500.org, https://www.top500.org/news/leaked-intel-roadmap-reveals-some-surprises-for-hpc-customers/

[100] October 30, 2018, https://globenewswire.com/news-release/2018/10/30/1639203/0/en/Supercomputing-Leader-Cray-Introduces-First-Exascale-class-Supercomputer.html

[101] October 30, 2018, http://investors.cray.com/phoenix.zhtml?c=98390&p=irol-newsArticle&ID=2374181

[102] November 12, 2018, https://globenewswire.com/news-release/2018/11/12/1649853/0/en/Atos-launches-new-BullSequana-Hybrid-supercomputer-for-AI-augmented-simulation.html

[103] November 29, 2018, https://www.nextplatform.com/2018/11/29/atos-rejiggers-sequana-supercomputers-adds-amd-rome-cpus/

[104] HPE, https://h20195.www2.hpe.com/v2/GetDocument.aspx?docname=a00016640enw&doctype=quickspecs&doclang=EN_US&searchquery=&cc=us&lc=en

[105] Ethernet Alliance, https://insidehpc.com/2018/03/new-2018-ethernet-roadmap-looks-future-speeds-1-6-terabits-s/

[106] Mellanox, http://www.mellanox.com/page/press_release_item?id=1899

[107] The Next Platform, https://www.nextplatform.com/2018/01/21/mellanox-trims-reach-profitable-1-billion/

[108] December 19, 2018, http://www.mellanox.com/page/products_dyn?product_family=280&mtag=sn3000_label

[109] Mellanox, http://www.mellanox.com/related-docs/prod_eth_switches/BR_SN3000_Series.pdf

[110] October 30, 2018, https://www.nextplatform.com/2018/10/30/cray-slingshots-back-into-hpc-interconnects-with-shasta-systems/

[111] October 30, 2018, https://www.cray.com/blog/meet-slingshot-an-innovative-interconnect-for-the-next-generation-of-supercomputers/

[112] Jeff Squyres, Atri Indiresay, Cisco, Meeting at SC18, Dallas, TX, US, 2018

[113] Matt Mohr, ARISTA, Meeting at SC18, Dallas, TX, US, 2018

[114] Dr. Niels Burkhardt, Extoll, Meeting at SC18, Dallas, TX, US, 2018

[115] Kurtis Bowman, Gen-Z booth, Meeting at SC18, Dallas, TX, US, 2018

[116] Paul Wade, Dolphin ICS, Meeting at SC18, Dallas, TX, US, 2018

[117] https://www.dolphinics.com/download/WHITEPAPERS/Dolphin_Express_reflective_memory.pdf

[118] https://spectralogic.com/2018/08/06/the-new-enterprise-drive-a-deeper-look-into-spectras-taos-feature-for-lto-tape-libraries/

[119] https://cds.cern.ch/record/2282014?ln=en

[120] https://indico.cern.ch/event/730908/contributions/3153156/attachments/1732268/2800425/LTO-CERN-HEPiX-Oct-2018-germancancio.pdf

[121] http://www.bdva.eu/sites/default/files/AI-Position-Statement-BDVA-Final-12112018.pdf

[122] https://www.prophetstor.com/federator-ai/

[123] November 6, 2018, https://insidehpc.com/2018/11/ddn-showcase-worlds-fastest-storage-sc18/

[124] November 13, 2018, https://www.ddn.com/press-releases/ddn-hpc-storage-innovation-leadership-recognized-hpcwire/

[125] January 22, 2019, https://www.ddn.com/press-releases/ddn-turnaround-tintri-wins-information-technology-deal-of-the-year-award/

[126] June 25, 2018, https://www.hpcwire.com/off-the-wire/storage-leader-ddn-acquires-lustre-file-system-capability-from-intel/

[127] January 30, 2019, https://www.ddn.com/blog/storage-industry-crystal-ball-2019-predictions/

[128] June 21, 2018, https://blog.seagate.com/enterprises/seagate-demonstrate-advanced-technology-ocp-summit-2018-support-accelerated-hyperscale-demand-data-growth/

[129] September 10, 2018, https://www.hpcwire.com/off-the-wire/seagate-unveils-industrys-most-advanced-14tb-data-storage-portfolio/

[130] September 11, 2018, https://insidehpc.com/2018/09/seagate-rolls-14-terabyte-hard-drives/

[131] August 9, 2018, https://blog.seagate.com/intelligent/open-data-center-summit-seagate-showcases-latest-enterprise-ssds-hdds/

[132] November 6, 2018, https://insidehpc.com/2018/11/predictions-sc18-change-climate-hpc/

[133] Consumer Technology Association - https://www.ces.tech/

[134] June 27, 2018, http://www.advancedclustering.com/technologies/beegfs-parallel-file-storage/

[135] October 26, 2018, https://www.hpcwire.com/off-the-wire/beegfs-based-burst-buffer-enables-world-record-hyperscale-data-distribution/

[136] November 14, 2018, http://www.advancedclustering.com/advanced-clustering-receives-beegfs-rising-bee-award-sc18/

[137] July 6, 2018, https://www.itnews.com.au/news/csiro-removes-hpc-bottleneck-with-storage-upgrade-496958

[138] November 26, 2018, https://www.hpcwire.com/2018/11/26/move-over-lustre-spectrum-scale-here-comes-beegfs/

[139] https://www.advancedhpc.com/solutions/parallel-file-systems/beegfs/

[140] Lubos Kolar, HPE, Meeting at SC18, Dallas, TX, US, 2018

[141] https://openpowerfoundation.org/

[142] S. Oehme, “Spectrum Scale – CORAL Enhancements”, Spectrum Scale User Group @ SC Asia, Singapore, March 2018 (http://files.gpfsug.org/presentations/2018/Singapore/Sven_Oehme_ESS_in_CORAL_project_update.pdf)

[143] S. Oehme, “Spectrum Scale performance update”, HEPIX 2016 (https://indico.cern.ch/event/531810/contributions/2306222/attachments/1357265/2053960/Spectrum_Scale-HEPIX_V1a.pdf).

[144] https://github.com/openstack/swiftonfile

[145] D. Hillebrand et al., “OpenStack SwiftOnFile: User Identity for Cross Protocol Access Demystified”, SNIA Storage Developer Conference, Santa Clara, 2015 (https://researcher.watson.ibm.com/researcher/files/us-dhildeb/OpenStack%20SwiftOnFile%20-%20User%20Identity%20for%20Cross%20Protocol%20Access%20Demystified.pdf)

[146] Fujitsu, “PRIMERGY CX400 M4”, 2017 (http://www.fujitsu.com/global/Images/primergy-cx400-m4.pdf).

[147] Fujitsu, “Fujitsu Completes Post-K Supercomputer CPU Prototype, Begins Functionality Trials”, June 2018 (http://www.fujitsu.com/global/about/resources/news/press-releases/2018/0621-01.html).

[148] Toshio Yoshida, “Fujitsu High Performance CPU for the Post-K Computer”, presentation at Hot Chips 2018.

[149] Yuichiro Ajima et al., “The Tofu Interconnect D”, IEEE Cluster 2018 (http://www.fujitsu.com/jp/Images/the-tofu-interconnect-d.pdf).

[150] Huawei, “Huawei Unveils Industry's Highest-Performance ARM-based CPU”, press release, 2019 (https://www.huawei.com/en/press-events/news/2019/1/huawei-unveils-highest-performance-arm-based-cpu).

[151] https://atos.net/en/products/high-performance-computing-hpc/bullsequana-x-supercomputers

[152] https://www.hpcwire.com/2018/08/14/cern-incorporates-ai-into-physics-based-simulations/

[153] https://www.nextplatform.com/2017/05/15/will-ai-replace-traditional-supercomputing-simulations/

[154] https://ldrd-annual.llnl.gov/ldrd-annual-2016/computing/brain

[155] http://www.teratec.eu/library/pdf/forum/2018/Presentations/A7_05_Alexis_Joly_Inria_Forum_Teratec_2018.pdf

[156] https://www.top500.org/news/julich-supercomputing-center-turns-to-modular-approach-to-deal-with-hpc-hpda-divide/

[157] S. Schmitt et al., “Neuromorphic Hardware In The Loop: Training a Deep Spiking Network on the BrainScaleS Wafer-Scale System”, 2017 International Joint Conference on Neural Networks (doi: 10.1109/IJCNN.2017.7966125).

[158] M. Davies et al., “Loihi: A Neuromorphic Manycore Processor with On-Chip Learning”, IEEE Micro, January/February 2018 (doi:10.1109/MM.2018.112130359).

[159] Intel, “Intel Creates Neuromorphic Research Community to Advance ‘Loihi’ Test Chip”, press release (https://newsroom.intel.com/editorials/intel-creates-neuromorphic-research-community)

[160] S. Furber et al., “The SpiNNaker Project”. Proceedings of the IEEE, 2014 (doi:10.1109/JPROC.2014.2304638).

[161] S. van Albada et al., “Performance Comparison of the Digital Neuromorphic Hardware SpiNNaker and the Neural Network Simulation Software NEST for a Full-Scale Cortical Microcircuit Model”, Front. Neurosci., 23 May 2018 (doi:10.3389/fnins.2018.00291).

[162] Jun Sawada et al., “TrueNorth Ecosystem for Brain-Inspired Computing: Scalable Systems, Software, and Applications,” SC16 Proceedings, 2016.

[163] PRACE-5IP D5.4 Deliverable “HPC Infrastructures Workshop #9”, 2018

List of Acronyms and Abbreviations

aisbl .....Association International Sans But Lucratif (legal form of the PRACE-RI)
AI.....Artificial Intelligence
ABCI .....Al Bridging Cloud Infrastructure
ASIC.....Application Specific Integrated Circuit
BDVA.....Big Data Value Association
BXI .....Bull eXascale Interconnect (product by ATOS)
CAGR.....Compound Annual Growth Rate
CoE.....Center of Excellence
cPPP .....contractual Public Private Partnership
CPU .....Central Processing Unit
CUDA.....Compute Unified Device Architecture (NVIDIA)
DoE.....(US) Department of Energy
DP.....Double Precision (64 bits)
EC.....European Commission
ECP.....Exascale Computing Project
ECI .....Exascale Computing Initiative
EFlop/s .....Exa (= 10^{18}) Floating point operations per second, also EFlops
EOSC.....European Open Science Cloud
ETP4HPC.....European Technology Platform for High Performance Computing
Fenix.....Federated ENgine for Information eXchange
FETHPC.....Generic term for HPC calls for projects and the series of related projects which have been selected, under H2020 FET programme (Future and Emerging Technologies), between 2014 and 2017
FP7 .....Framework Programme 7
FPA.....Framework Partnership Agreement
FPGA.....Field Programmable Gate Array
GB .....Giga (= 2^{30} ~ 10^9) Bytes (= 8 bits), also GByte
Gb/s .....Giga (= 10^9) bits per second, also Gbit/s
GB/s.....Giga (= 10^9) Bytes (= 8 bits) per second, also GByte/s
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 (= 10^9) Floating point operations (usually in 64-bit, i.e. DP) per second, also GF/s
GHz .....Giga (= 10^9) Hertz, frequency =10^9 periods or clock cycles per second
GPU.....Graphic Processing Unit
GT/s.....Giga (10^9) transfers per second
HBM.....High Bandwidth Memory
HBP .....Human Brain Project
HDR .....InfiniBand interconnect generation called High Data Rate with link speed 200Gb/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
HPDA .....High-Performance Data Analytics

HPL .....High-Performance LINPACK
ICEI .....Interactive Computing E-Infrastructure for the Human Brain Project
ISC.....International Supercomputing Conference; European equivalent to the US based SCxx conference. Held annually in Germany.
JU .....Joint Undertaking
KB .....Kilo (= 2^{10} ~10^{3}) Bytes (= 8 bits), also Kbyte
LINPACK.....Software library for Linear Algebra
LTO .....Linear Tape-Open
MB.....Management Board (highest decision making body of the PRACE project)
MB.....Mega (= 2^{20} ~ 10^{6}) Bytes (= 8 bits), also MByte
MB/s .....Mega (= 10^{6}) Bytes (= 8 bits) per second, also MByte/s
MFlop/s .....Mega (= 10^{6}) Floating point operations (usually in 64-bit, i.e. DP) per second, also MF/s
ML.....Machine Learning
MPI.....Message Passing Interface
NIC .....Network Interface Controller
NSCI.....National Strategic Computing Initiative (USA)
NUDT.....National University of Defence Technology (China)
NUMA.....Non-Uniform Memory Access
NVMe.....Non-Volatile Memory Express
OPA.....Omni-Path Architecture
PAM-4.....Pulse Amplitude Modulation signalling for high-speed serial links
PCP.....Pre-Commercial Procurement
PFlop/s.....Peta (= 10^{15}) Floating-point operations (usually in 64-bit, i.e. DP) per second, also PF/s or PF
PPI.....Public Procurement of Innovative solutions
PRACE.....Partnership for Advanced Computing in Europe; Project Acronym
PRACE 2.....The current phase of the PRACE Research Infrastructure following the initial five-year period.
QSFP .....Quad Small Form-factor Pluggable
R&D .....Research and Development
R&I.....Research and Innovation
RI.....Research Infrastructure
RIA .....Research and innovation action (type of H2020 project)
SGA.....Specific Grant Agreement
SMT.....Simultaneous Multithreading
SRIA.....Strategic Research and Innovation Agenda
SVE .....Scalable Vector Extension
TB.....Technical Board (group of PRACE Work Package leaders)
TB.....Tera (= 2^{40} ~ 10^{12}) Bytes (= 8 bits), also TByte
TCO.....Total Cost of Ownership. Includes recurring costs (e.g. personnel, power, cooling, maintenance) in addition to the purchase and set-up cost.
TFlop/s .....Tera (= 10^{12}) Floating-point operations (usually in 64-bit, i.e. DP) per second, also TF/s or TF

TOp/s.....Tera operations per second, also TOps or Tops or Top/s
Tier-0.....Denotes the apex of a conceptual pyramid of HPC systems. In this context, the Supercomputing Research Infrastructure would host the Tier-0 systems; national or topical HPC centers would constitute Tier-1
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 Intensif, 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 second deliverable of PRACE-5IP Work Package 5 “Task 5.1 - Technology and market watch” and represents 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.

The Top500 (Nov 2018) list is still dominated by China and the United States, that rank very close to each other, with, this time, the United States in the first and second place, with 21.8% of system share and 37.7% of performance share, China on third and fourth position, with 45.4% of system share and 31% of performance share. Two machines installed in Europe are in the Top10: SuperMUC-NG (LRZ, Germany) and Piz Daint (CSCS, Switzerland). Japan still dominates the Green500 list with four supercomputers ranked in the Top10. The first and second ranked supercomputers in the Top500 list, Summit and Sierra, are also the machines showing the best HPCG performance. The only European supercomputer in the Top10 of HPCG is Piz Daint. A new benchmark related to IO performance (IO500) initiated in 2017 is gaining momentum with 63 entries while a proposal for a new benchmark on large scale deep learning called Deep500 have been presented.

Plans for exascale are well-defined in China, USA and Japan. In China, three tracks are explored in parallel with three prototypes deployed by Sugon, Tianhe, and Sunway (ShenWei) in 2018. In Japan, the post-K project, targeting exascale class supercomputer in 2021, is progressing on schedule with the test started on the first prototype of the ARM-based CPU developed by Fujitsu. In the United States, the ECP project is addressing application development, software technology, and hardware technology and exascale systems testbeds. In Europe, EuroHPC is a Joint Undertaking established in 2018 with, as one of its goals, to construct an exascale supercomputer based on European technology.

HPC as Cloud computing offering has already been available for some time and is a real option for some HPC workloads. It generally complements, rather than replaces, the traditional HPC systems. Big players like Amazon, Microsoft and Google have commercial offerings that target the HPC market. The main trend in 2018 for HPC is bare metal service, while solutions based on containers have some advantages in terms of redundancy and enhanced information security. Offers including GPU, targeting mostly AI, may also be of interest for HPC. Even quantum computing (IBM Quantum System for example) is also available in the cloud.

Regarding the consolidation in the HPC market, the acquisition of Cavium by Marvell is an important event since the ThunderX ARM-based processors are among the most promising for HPC.

In the EU landscape, EuroHPC is the central and overarching new piece, consolidating or recomposing the ecosystem. It brings the promise of better coordination and more funding for HPC in Europe, both for infrastructures and for R&I, including applications. In this context, existing entities like PRACE, ETP4HPC, and BDVA, need to evolve in order to take into account this new context. Among the coordination and support actions, EXDCI-2 continues the coordination of the

HPC ecosystem with important enhancements with respect to EXDCI, so as to better address the convergence of big data, cloud and HPC. On the R&I side, centers of excellence and FET HPC projects remain important and active players in the field while the new EPI (European Processor Initiative) is of major importance for Europe. On the infrastructure side, PPI4HPC and HBP are contributing respectively for deploying innovative equipment in HPC centers and developing a data infrastructure complementing the current supercomputer infrastructure.

In terms of core technologies for HPC system processors, Intel is now facing the competition of new competitors: AMD with the EPYC architecture, and Marvell, with the ThunderX ARM-based processor family. Several large systems based on these processors have already been announced. The ARM ecosystem is very active with processors targeting HPC announced by Huawei and Fujitsu. The POWER processor, coupled with NVIDIA GPU, is used for the supercomputers Summit and Sierra, which are listed on position #1 and #2 of the Top500 list as of November 2018. In terms of accelerators for HPC, these systems mostly rely on NVIDIA GPUs with AMD being so far the only competitor but with less success than NVIDIA. FPGA is still a niche market.

The most used volatile memory is currently DDR4 which tops with a DIMM size of 128GB running at 2933MT/s. Another class of volatile memory technology important to HPC is High Bandwidth Memory (HBM), today usually implemented as stacked memory, allowing parallel access to multiple (up to 8) slices. It supports speeds up to 2GT/s. Non-volatile memory technologies are still emerging but may be used in the future on compute node and on IO systems.

Regarding data storage and data management, technologies and components are evolving, new services are being deployed some targeting specific workload like AI. Tape storage is still widely used with capacity increasing over time. This is the same for disk storage, DDN being the main player. Flash and non-volatile memories are expected to play an increasing role in the near future.

In terms of vendor solutions and roadmaps, HPE remains in the 2nd position in the HPC world with products derived from SGI (acquired by HPE). HPE has installed the largest ARM-based cluster in Sandia National Lab. Fujitsu continues to be the main provider of high-end HPC solutions in Japan. Chinese vendors (mostly Huawei, Lenovo) are actively investing in the HPC market. Atos/Bull currently has 22 supercomputers ranked in the Top500 and has announced a blade for ARM processors.

Paradigm shifts in HPC technologies include AI, heterogenous architectures, and neuromorphic computing. Since our last report, AI is gaining momentum in a number of dimensions: to assist regular HPC simulations (“augmented HPC”) or to replace HPC simulation. HPC can also be used to create high fidelity training AI databases. Modular heterogeneous solutions are becoming widespread with the double objectives of limited power consumption and programmability. In this approach, the user will no longer see a supercomputer as a single homogeneous entity but rather as an aggregation of specialized modules sharing common resources such as parallel file systems or visualization nodes on a central network. The term “neuromorphic computing” broadly refers to compute architectures that are inspired by features of brains as found in nature. There is a strong interest, at the level of research, into such devices in the context of brain modelling as well as artificial intelligence and machine learning.

1 Introduction

The PRACE-5IP Work Package 5 (WP5), “HPC Commissioning and Prototyping”, is defined by three main tasks: T5.1 "Technology and market watch, vendor relationships independent of procurements"; T5.2 "Best practice for energy-efficient HPC center infrastructures design and operations"; T5.3 "Extended best practice guide for prototypes or demonstrators". WP5 aims to deliver updated information on the HPC ecosystem, market surveys and best practices, providing guidance to decision-makers at all levels.

In particular, Task 5.1 focuses on assessing the match between the current and upcoming system architectures and the requirements of the users. To this end, a strong effort was devoted to tracking the evolution of HPC technologies based on vendor roadmaps, as well as collecting updates coming from EU projects and initiatives. This task is the continuation of a well-established effort, using assessments of the HPC market based on market surveys, Top500 and Green500/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. The outcome of these activities is reported in this deliverable, that builds on the important work performed for the previous document – “Market and Technology Watch Report Year 1” – and complements its overall view.

The deliverable contains many technical details on recent HPC topics. It 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 this may affect planning for future data centers and systems. Users should refer to this document as an overall view of HPC hardware and expect some of the solutions described to be available to them in the near future. Developers should read this to gain knowledge of recent and future trends, in order to make informed decisions on the development of their applications.

The deliverable is organized into six main chapters. Besides this introduction, Chapter 1, it contains:

- Chapter 2: “Worldwide HPC landscape and market overview”, which provides a geographical and market analysis, an overview of large EU and worldwide initiatives, and current trends on HPC cloud computing;
- Chapter 3: "EU landscape overview" which focuses on EU landscape, providing an update on the main projects and technology oriented initiatives;
- Chapter 4: “Core technologies and components” is a quick overview of processors, accelerators, memory and storage technologies and interconnect technologies;
- Chapter 5: "Data Storage and Data Management – Technologies and Components" provides an overview of present storage models, architectures and solutions;
- Chapter 6: "Overview of Vendor solutions/roadmaps" tracks the evolution of main players’ roadmaps;

- Chapter 7: “Paradigm shifts in HPC technologies” which reports the most promising technologies for future HPC systems, including neuromorphic, AI and heterogeneous architectures.

2 Worldwide HPC landscape and Market overview

2.1 Quick update 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], the HPCG [3] and the IO500 [4] benchmarks. 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 2018 results for R_{max} (HPL performance) values in Top500 (TFlop/s), power efficiency (GFlop/s/watt) values in Green 500, HPCG (PFlop/s) and IO500 ($\sqrt{GiB * IOP/s}$) are presented Table 1.

	Top500	Green500	HPCG	IO500
1	USA (Summit)	Japan (Shoubu system B)	USA (Summit)	USA (Summit)
2	USA (Sierra)	USA (DGX SaturnV Volta)	USA (Sierra)	UK (Data Accelerator)[a]
3	China (Sunway TaihuLight)	USA (Summit)	Japan (K Computer)	Republic of Korea (Nurion)
4	China (Tianhe-2A)	Japan (ABCI)	USA (Trinity)	Japan (Oakforest-PACS)[b]
5	Switzerland (Piz Daint)	Japan (TSUBAME3.0)	Japan (ABCI)	USA (WekaIO)
6	USA (Trinity)	USA (Sierra)	Switzerland (Piz Daint)	Saudi Arabia (ShaheenII)[c]
7	Japan (ABCI)	Japan (AIST AI Cloud)	China (Sunway TaihuLight)	UK (Data Accelerator)[d]
8	Germany (SuperMUC-NG)	Spain (MareNostrum P9 CTE)	Republic of Korea (Nurion)	USA (Exascaler on GCP)
9	USA (Titan)	China (PreE)	Japan (Oakforest-PACS)	Japan (Oakforest-PACS)[e]
10	USA (Sequoia)	Taiwan (Taiwania 2)	USA (Cori)	Saudi Arabia (ShaheenII)[f]

[a] Filesystem: Lustre, Client nodes: 512, Client total procs: 4224

[b] Filesystem: IME, Client nodes: 2048, Client total procs: 16384[c] Filesystem: DataWarp, Client nodes: 1024, Client total procs: 8192

[d] Filesystem: BeeGFS, Client nodes: 184, Client total procs: 5888

[e] Filesystem: Lustre, Client nodes: 256, Client total procs: 8192

[f] Filesystem: Lustre, Client nodes: 1000, Client total procs: 16000

Table 1: Top10 systems in benchmark results for Top500/Green500/HPCG/IO500.

Following the sudden increase in 2015, China caught up to the USA in terms of R_{max} values. In 2017, both countries shared almost the same R_{max}. However, starting from the June 2018 list of the Top500, USA again took the leadership (approximately 7% according to cumulative R_{max} values of respective countries) from China with the help of the most recent systems called Summit and Sierra. In 2018, Germany has managed to enter the Top10 of the Top500 list with the most recent SuperMUC-NG machine. With this "attack" of Germany, a second machine from Europe (the other

being Piz Daint) has entered the Top10 of Top500. In the Top10, there is also a new supercomputer from Japan (AI Bridging Cloud Infrastructure (ABCI)) specially designed for artificial intelligence, machine learning, and deep learning.

Japan still dominates the Green500 list with four supercomputers ranked in the Top10, including the top position. Similar to the Top500, the impact of USA increases in the Green500, with three computers in the Top10 list. Finally, the last three slots in the Top10 of the Green500 are taken by Spain, China and Taiwan.

The first and second-ranked supercomputers in the Top500 list, Summit and Sierra, are also the machines showing the best HPCG performance. The only European supercomputer in the Top10 of HPCG is Piz Daint. Mostly USA and Japan supercomputers dominate the HPCG list. It is also worth to mention that a supercomputer (Nurion) from the Republic of Korea appears in the HPCG list.

Computational performance is the most significant metric used to determine the fastest supercomputer and generally the effect of I/O is neglected. In order to track the performance storage systems, the IO500 [4] benchmark was initiated in mid-2017. Although the IO500 list aims to include the storage performances of all supercomputers listed in the Top500, currently (November 2018) the list comprises only 63 entries. The new giant supercomputer of the USA: Summit (storage vendor: IBM and filesystem: Spectrum Scale), takes the first rank of the IO500. Summit is followed by UK's supercomputer, Data Accelerator (storage vendor: Dell EMC and filesystem: Lustre). As indicated in Table 1, the IO500 list may include benchmarking results for a specific supercomputer more than once. In particular, by using different storage filesystems, a different number of nodes and processors, different IO500 scores are calculated for the same supercomputer. For example, the Data Accelerator system hosted in University of Cambridge has two entries in the Top10 of IO500: the first one (rank 2) uses Lustre with 512 nodes and the second one (rank 7) uses BeeGFS with 184 nodes. Finally, its worthy to notice the inclusion at #5 of Weka.IO, a new US company providing a flash native parallel object file system called WekaIO Matrix.

Figure 1 and Figure 2 present the system and performance share of the countries according to the November 2018 Top500 benchmarks. In these figures, the system share of a country is presented by the number of systems present and the performance share by the total R_{max} values. Since the Green500 list includes the same systems as the Top500 in a different order, the system share results for the Green500 are identical to the Top500.

In terms of the number of systems in Top500 and Green500 (see Figure 1), China covers almost half of the systems (45.4%). Approximately, a quarter of the Top500 supercomputers are hosted in the USA. The share of Japan is only 6%. The quota of Europe in the system share is around 20% and the major contributions are coming from the UK (4%), France (3.6%) and Germany (3.4%).

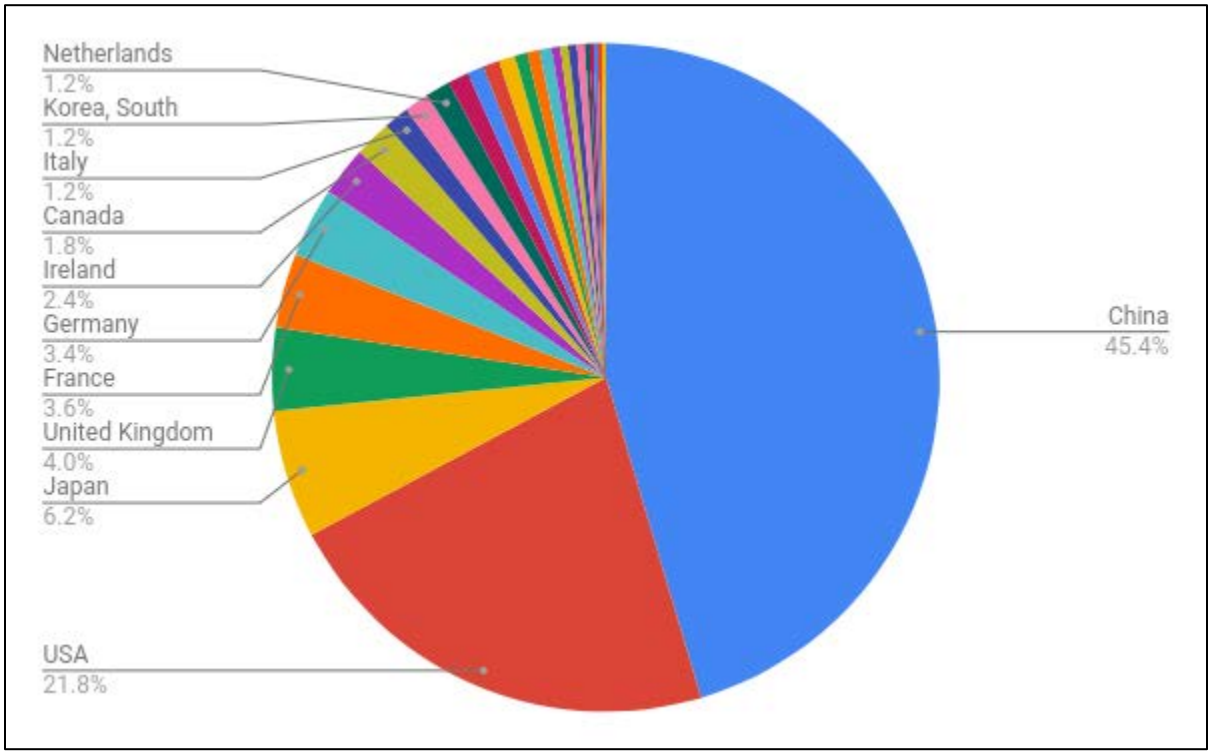

Figure 1: System share in Top500 and Green500.

As shown in Figure 2, even though China has half of the Top500 supercomputers, China's performance (R_{max}) share is only 31%. With the amazing contribution of Summit (143 PFlop/s) and Sierra (95 PFlop/s), USA offers nearly 38% of the performance share of the Top500. China is followed by Japan by 7% performance share. In contrast to system share values shown in Figure 1, Germany provides the biggest contribution (4.3%) from Europe and followed by France (3.1%) and the UK (2.9%). Although Switzerland does not have so many supercomputers, Switzerland comprises 1.6% of the performance share of the Top500 mostly due to the contribution of Piz Daint (21 PFlop/s). South Korea's performance share is the same as Switzerland's.

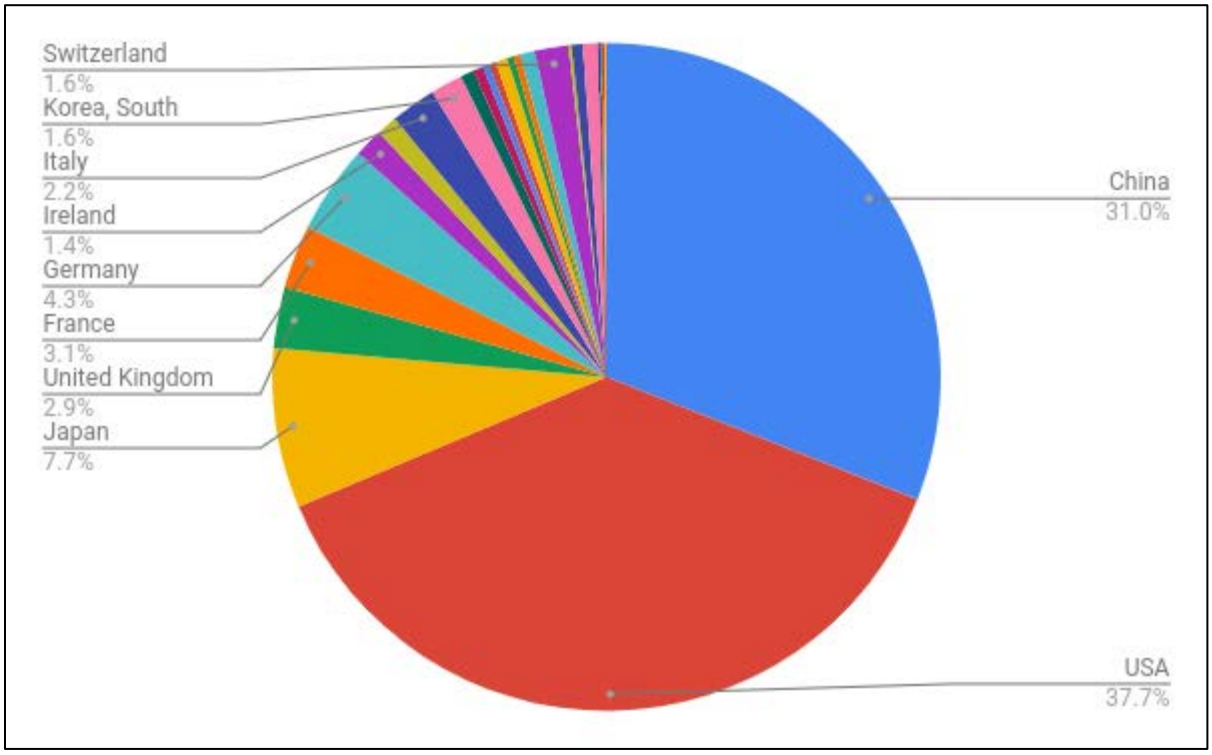

Figure 2: Performance share in Top500.

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. Figure 3 shows distinctly that the system share trends of the USA and China are the opposite of each other. In particular, when China boosts the number of systems in 2015 from 8 to 34%, a reduction starts in the USA from 47 to 34%. Then, both USA and China equally share the number of systems in between mid of 2016 and 2017. This tie is broken with the second attack of China from 34 to 46%. During this second "attack" of China, another downward trend in the USA has been observed and the system share of the USA was reduced to 22%. Summit is followed by UK's supercomputer, Data Accelerator (storage vendor: Dell EMC and filesystem: Lustre).

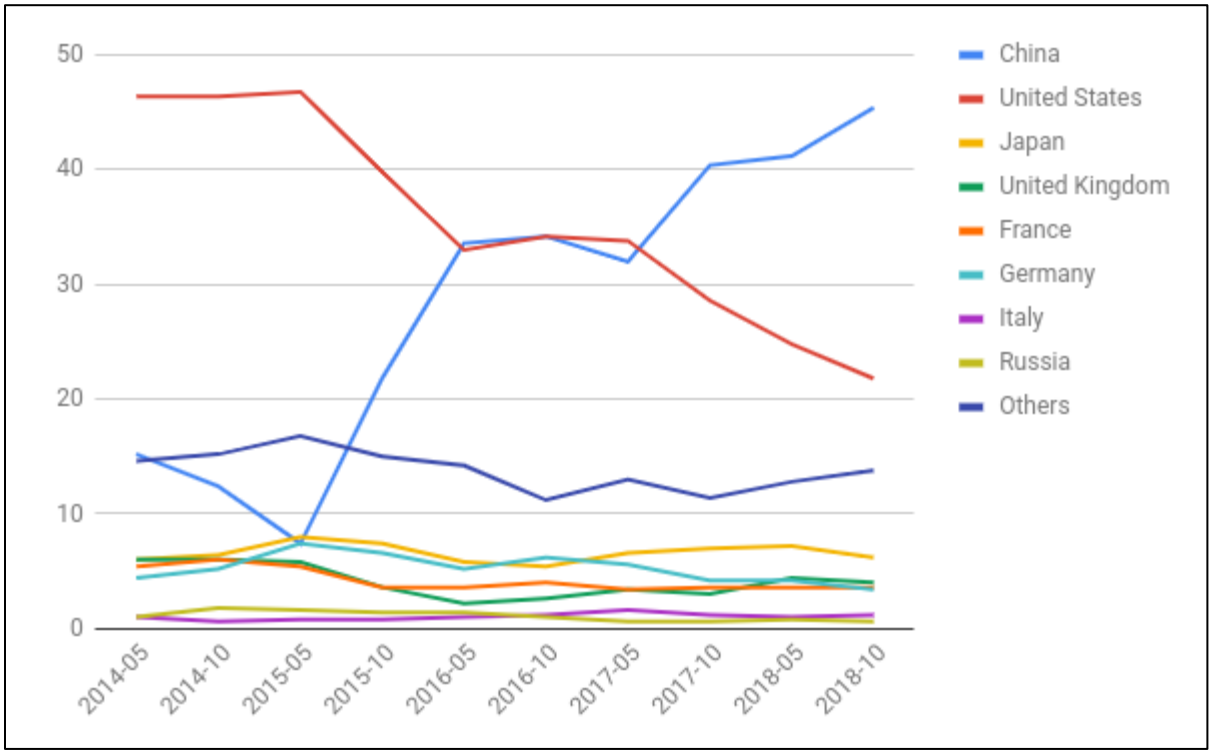

Figure 3: Countries system share over time (Top500).

Figure 4 shows the percentage of cumulative R_{max} values for countries listed in Top500. In parallel to the increase in the number of systems, China approaches the USA and even exceeds slightly in the mid of 2016. However, in the mid of 2018, the picture is reversed in favor of the USA due to the recent installations (Summit and Sierra). The gradual increase of Japan (10.5%) until the end of 2017 has continuously started to decline, falling to 7.7% at the end of 2018. There is no significant change in the status of European and other countries.

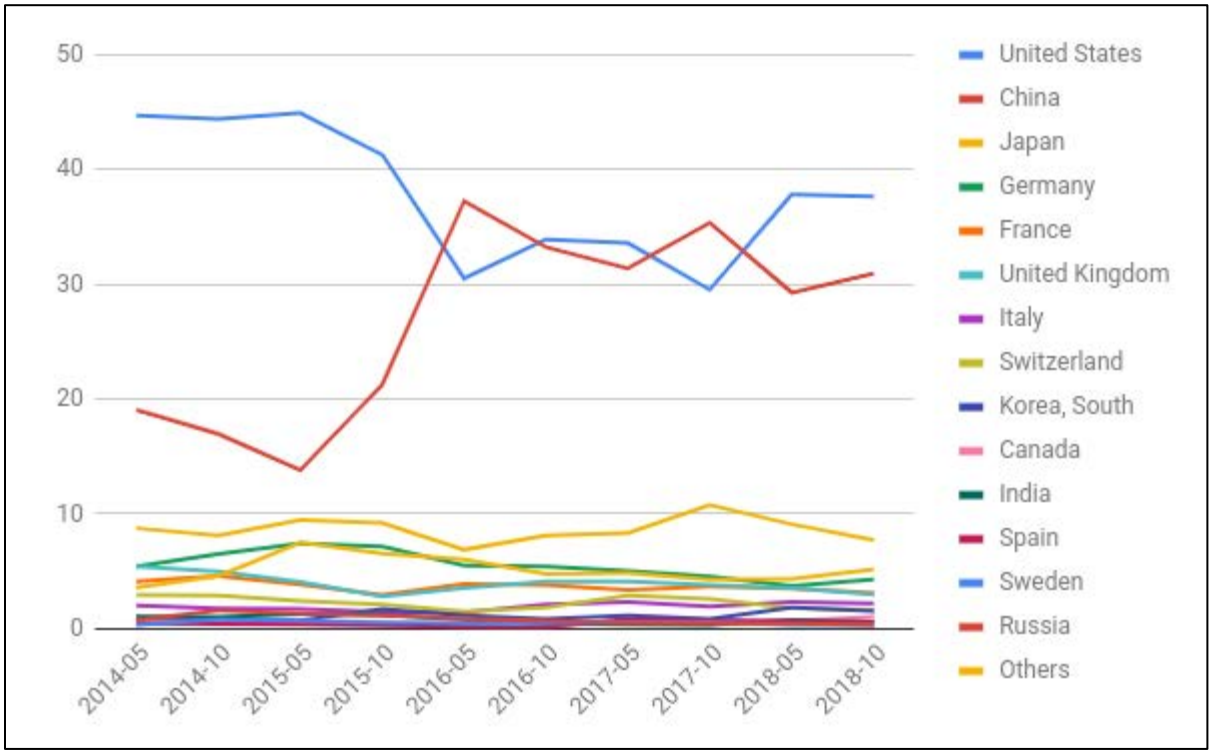

Figure 4: Percentage of cumulative R_{max} values (in GFlop/s) for countries (Top500). Y-axis represents the percentage of R_{max} values.

Table 2 shows the number of systems operating in the leading countries. The data included in this table are taken from 2016 to 2018 Top500 lists. It is apparent that the number of systems in China grows year by year and currently is twice the number of the US systems. The increasing trend is also visible for the UK. On the other hand, the number of systems in Japan and France are largely stable, but the same metric for Germany is in decline. In particular, in 2018, the number of systems is halved for Germany compared to 2016.

Top500 Systems	China	USA	Japan	UK	France	Germany
2016-11	171	171	27	13	20	31
2017-11	202	143	35	15	18	21
2018-11	227	109	31	20	18	17

Table 2: Leading countries systems shares in the Top500.

Figure 5 highlights the percentage of cumulative R_{max} values for the European countries. Almost in all countries, the decline in performance share is obvious. Germany holds the leadership and France and the UK are the followers.

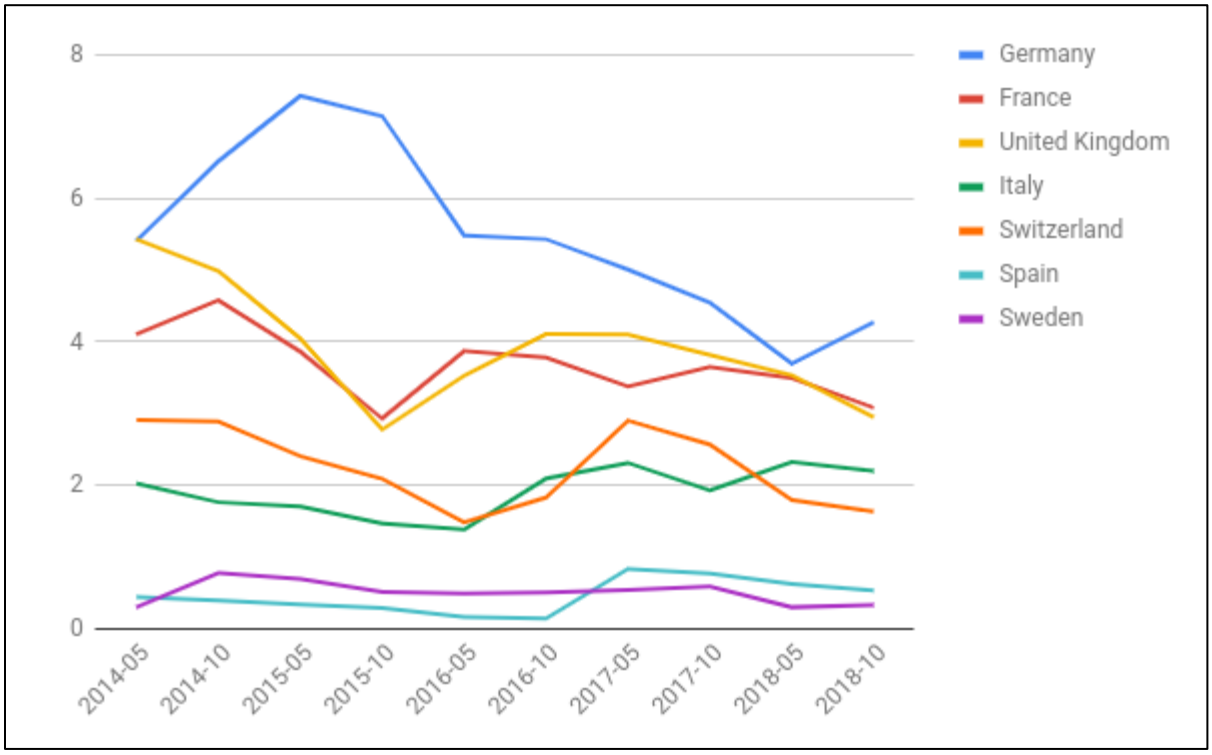

Figure 5: Percentage of cumulative R_{max} values (in GFlop/s) for European countries.

Figure 6 and Figure 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 to reflect percentage compared to the size of the corresponding list. In 2017, there was only one supercomputer (Piz Daint) in top 10 of the Top500 list. With the help of Germany's new machine (SuperMUC-NG), the total number of European supercomputers in Top10 increased to two in 2018. There are 5 and 15 machines in Top20 and Top50 lists of 2018, respectively. These values indicate a slight decrease compared to 2014.

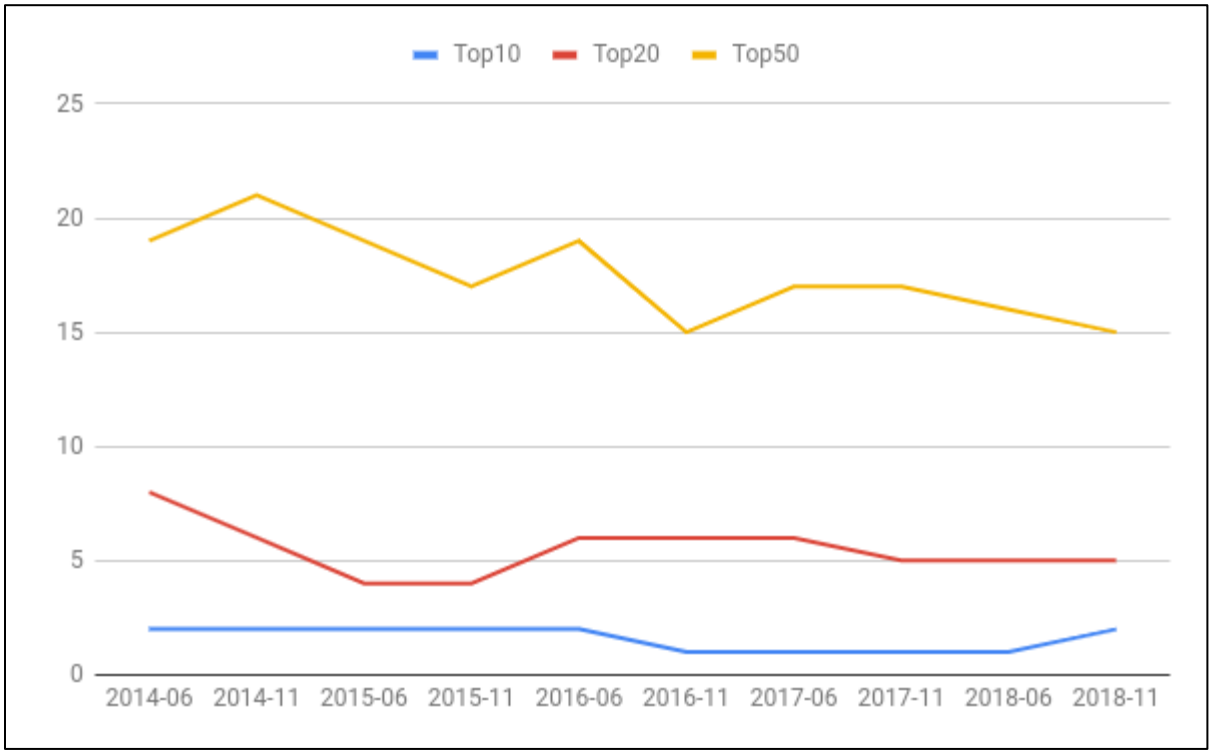

Figure 6: Systems share in Top10/20/50 for Europe.

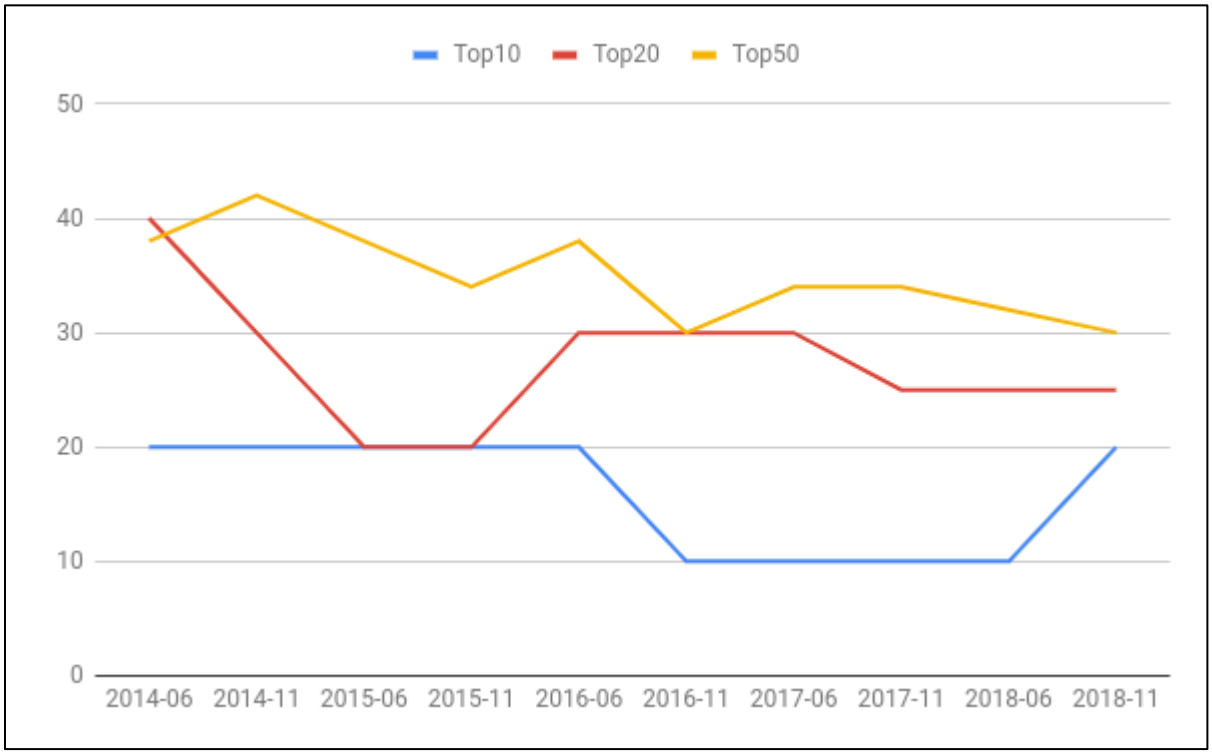

Figure 7: Percentage of systems in Top10/20/50 for the Europe countries.

2.1.2 Accelerators

Figure 8 shows the fraction of systems equipped with accelerators in the Top50. This figure indicates that more than half of the Top50 (30) systems include an accelerator, with the NVIDIA GPU being the preferred choice. Since Intel has withdrawn the production of Xeon Phi processors, its existence as an accelerator is not visible anymore. It should also be noted that the portion of Intel Xeon Phi co-processor equipped systems was continuously decreasing since 2015 and the gap left behind by it has been filled by NVIDIA. Deep computing processor and Matrix-2000 (developed by Chinese National University of Defense – NUDT) accelerators are the newcomers of 2018 and four HPC systems are powered by these processors.

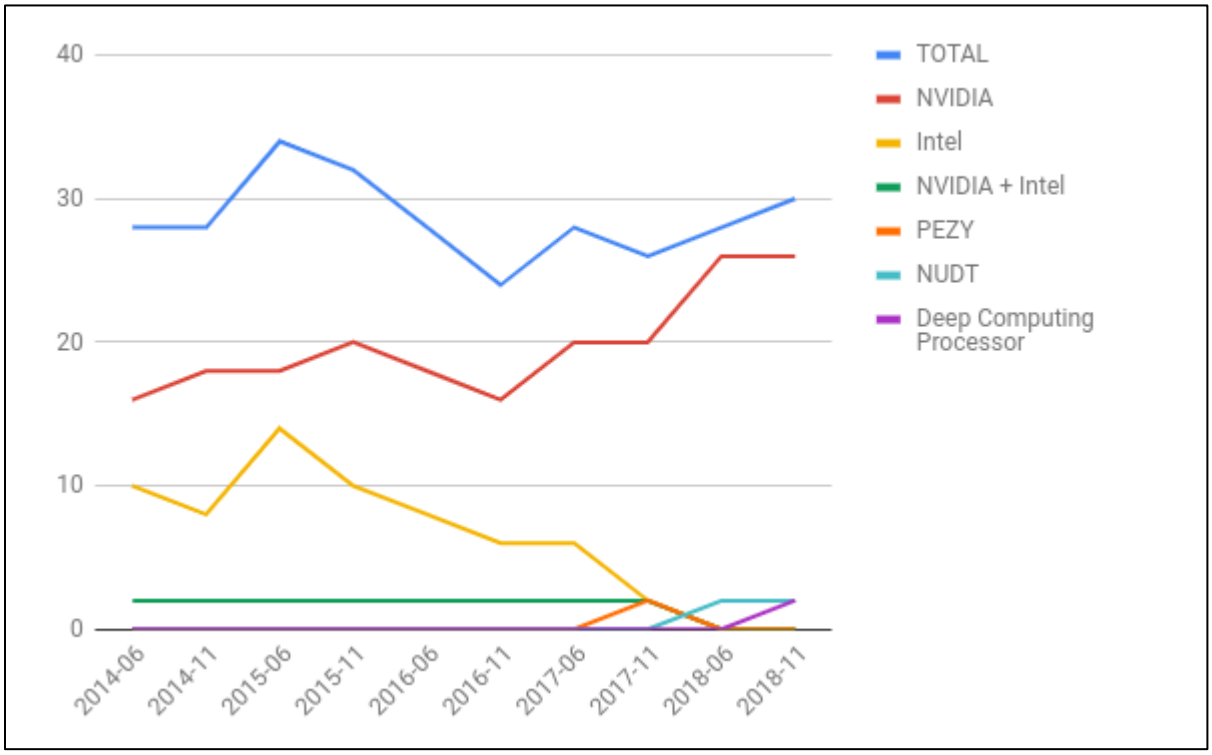

Figure 8: Fraction of systems equipped with accelerators (Top 50).

Figure 9 compares the fraction of accelerators in both Europe and the world based on Top50 and Top500 rankings (Euro50 means systems based in the Europe part of the 50 first systems in the Top500; Euro500 means systems based in the Europe part of the Top500 list). It is clear that NVIDIA GPU dominates both November 2018 Top50 and Top500 lists. Figure 9 also shows that 28% of the systems in the Top500 are equipped with an accelerator. Compared to November 2017, the current values indicate an additional 8% increase in the number of systems powered by accelerators in the world. Similarly, to the global picture, the number of accelerated systems in the European countries is also in the fast increase trend: in Euro50 (from 6 to 20%) and in Euro500 (from 2% to 12%) compared to last year's statistics.

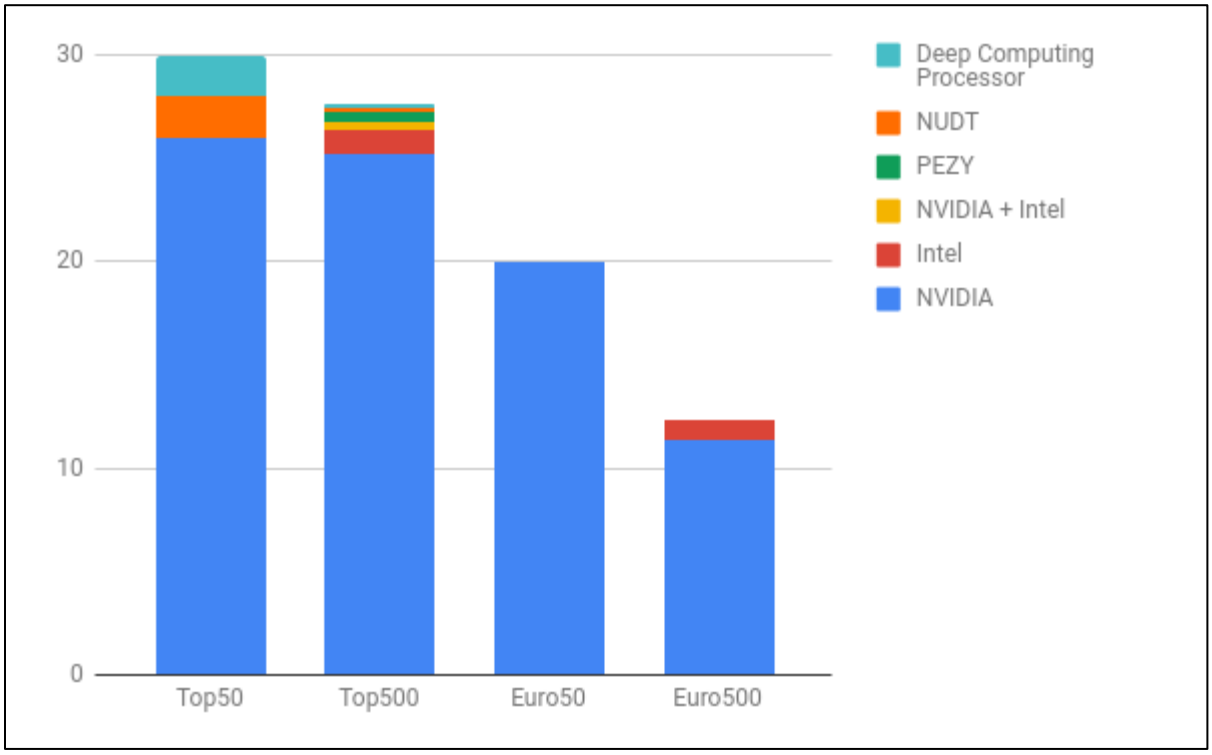

Figure 9: Fraction of systems equipped with accelerators (November 2018).

2.1.3 Age

Figure 10 shows the average age of systems (in terms of time of presence in the Top500) globally and for Europe. It is apparent that the average age of systems oscillates between 1 and 2 years for the last 5 years for the Top50 and Top500 for both Europe and the world.

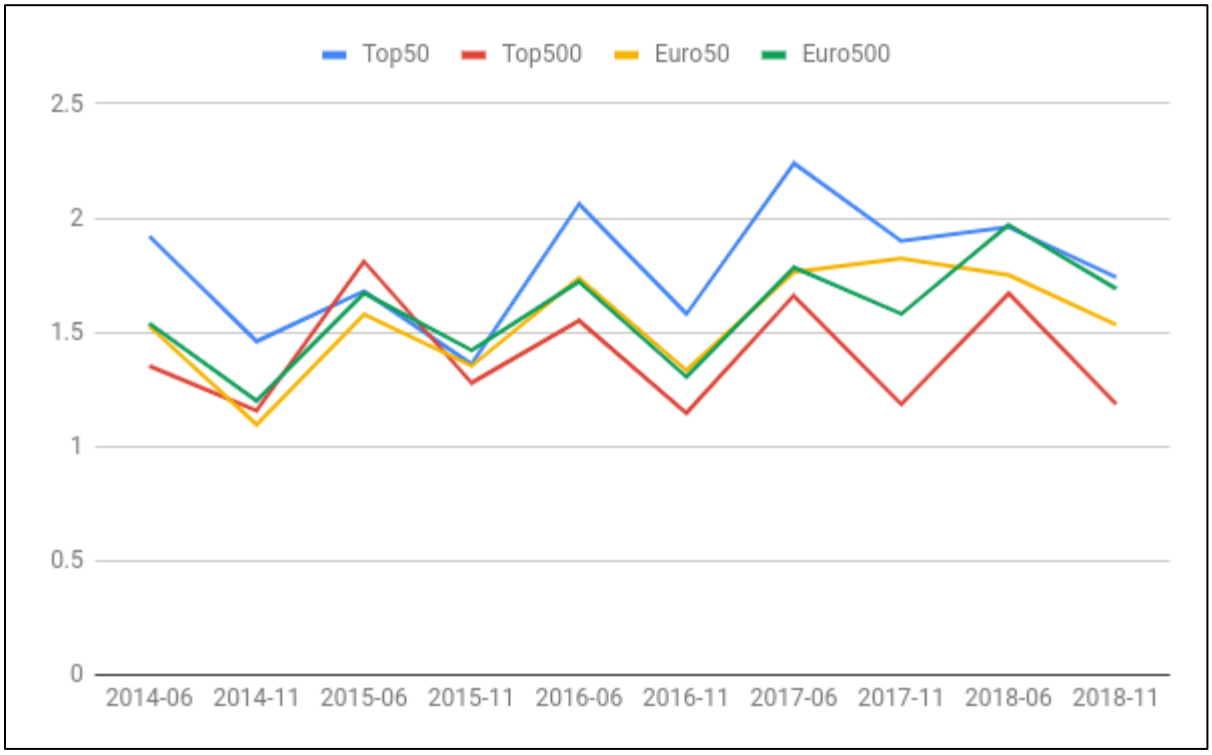

Figure 10: Average age of the systems in years.

2.1.4 Vendors

Figure 11 and Figure 12 show the position of vendors in the Top50, globally and local to Europe. Even though Cray is still the dominant vendor with 16 systems in the world, Cray tends to fall starting from the end of 2016. HPE is in a strong attack starting in the same year as the fall of Cray has been observed, which places HPE to the second rank in the Top50. The acquisition of SGI by HPE in 2016 may also have been a major contributor in this upward trend. The number of IBM systems, on the other hand, has been steadily decreasing especially starting from 2013, and this decrease cannot be attributed alone to the sale of IBM x86 business to LENOVO in 2014. For the other vendors, there is no significant change regarding their positions in Top50. The picture seen in Figure 11 is to some extent also valid for Europe (see Figure 12).

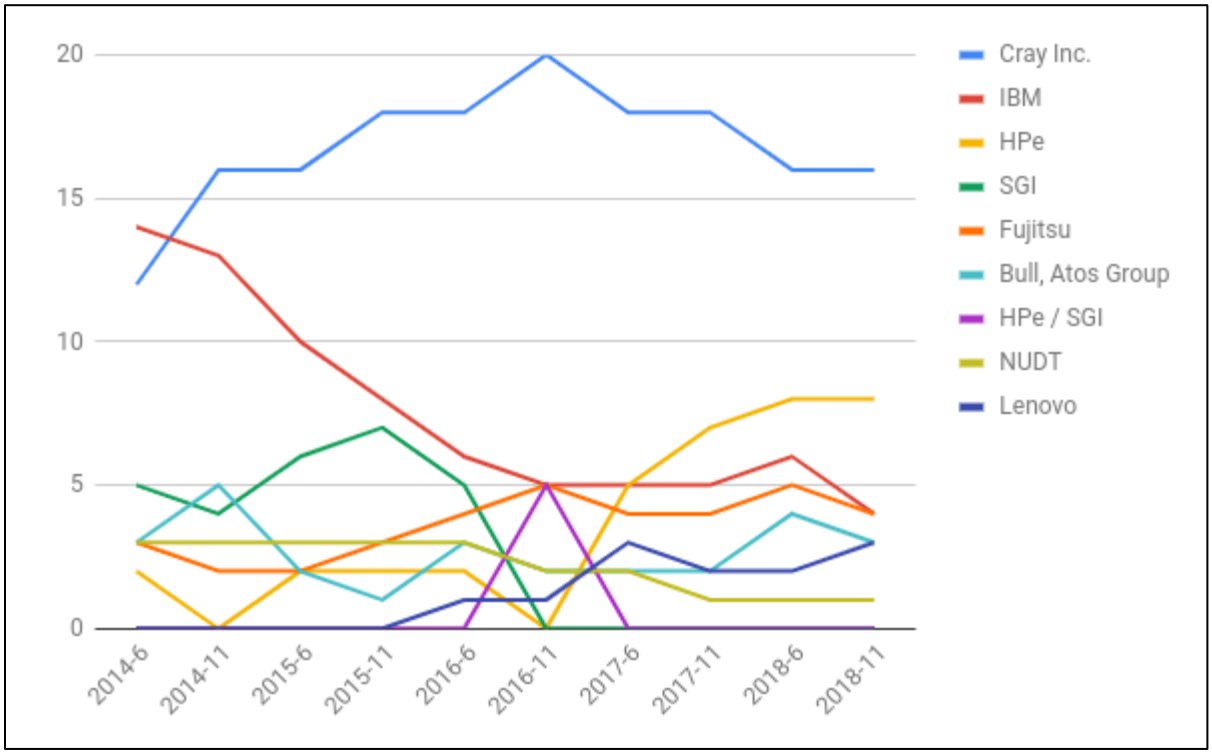

Figure 11: Top50 vendors (world).

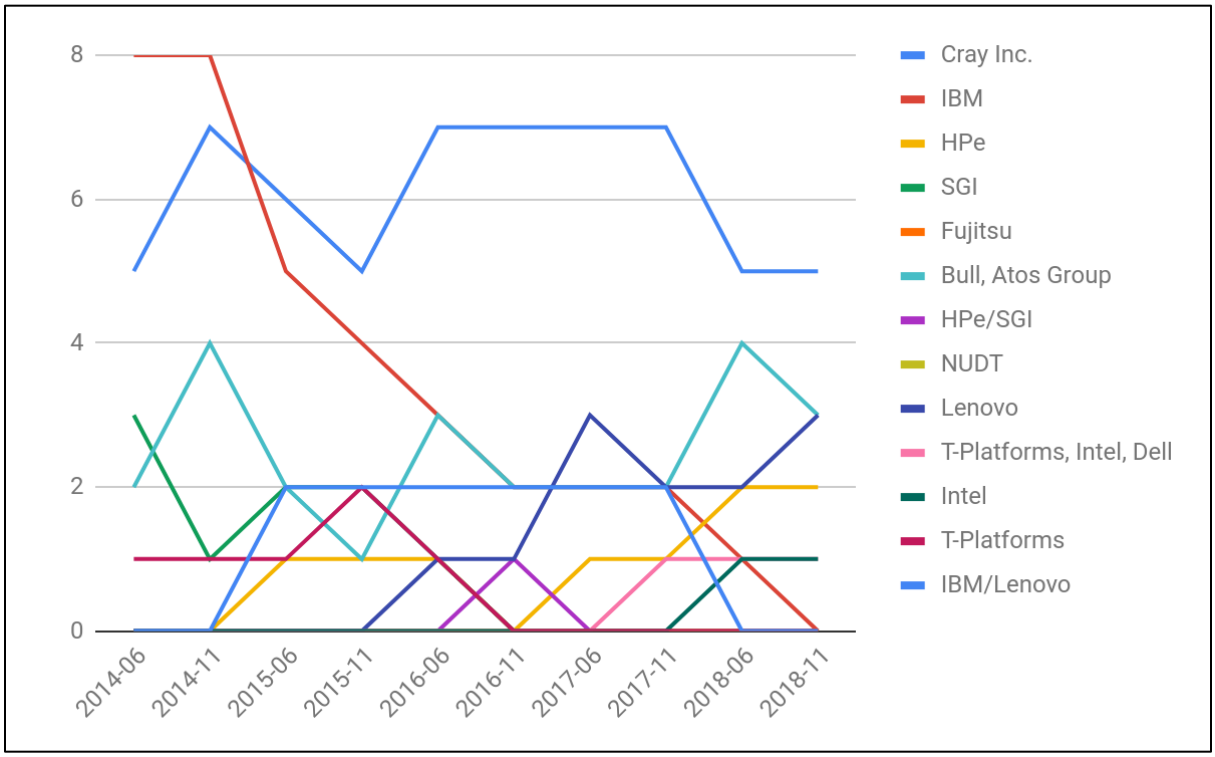

Figure 12: Top50 vendors (Europe).

Figure 13 shows the performance of the European vendor Bull. Most of the Bull systems are installed in Europe, and the number of Bull systems shows a wavy motion especially in the Top50

list. In the Top100, the number of Bull powered machines has increased between 2014 and 2015. Currently, in the Top500 list, 23 HPC systems were provided by Bull. As shown in Figure 14, the highest-ranking system from Bull was ranked 43rd in the Top500 in 2015, marking its lowest point in recent years. Following this, Bull starts to climb the summit and now the HPC machine at the rank 16 is powered by Bull. The other vendors from Europe, MEGWARE and ClusterVision, have much lower representation in the Top500 lists. In November 2018, each of these vendors had two systems in Top500. Additionally, ClusterVision has declared bankruptcy in 2019.

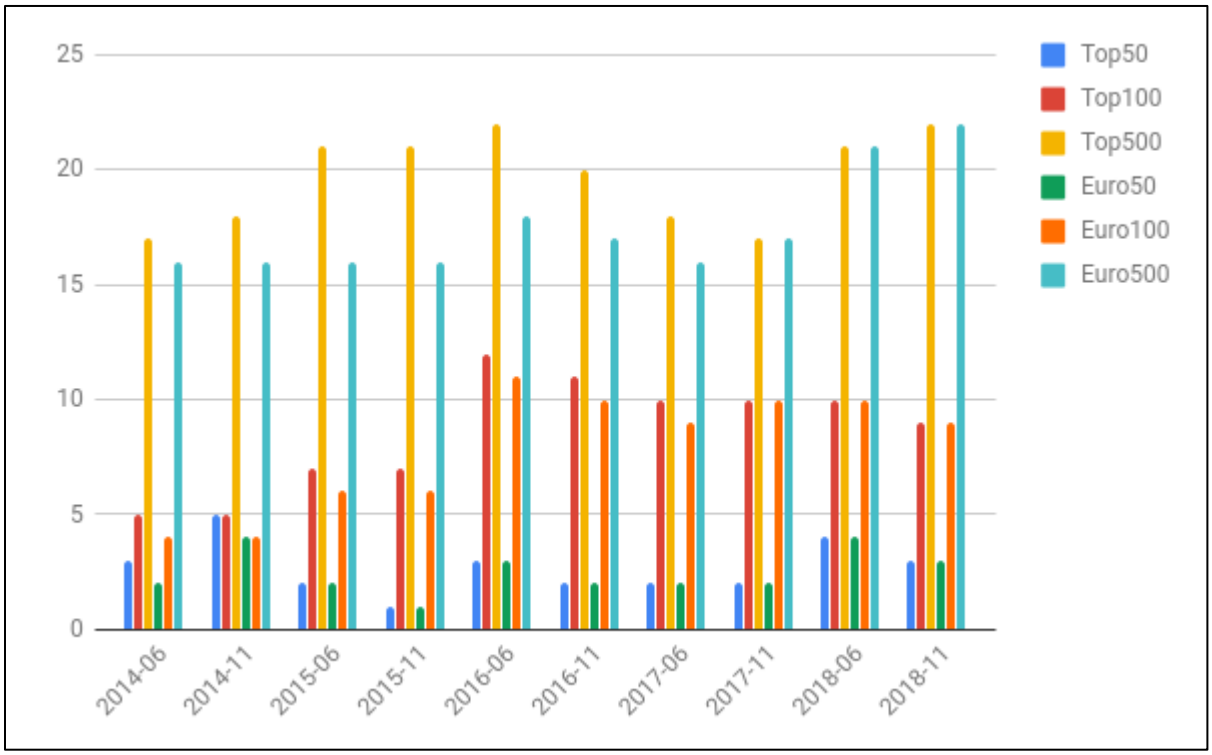

Figure 13: The number of Bull systems.

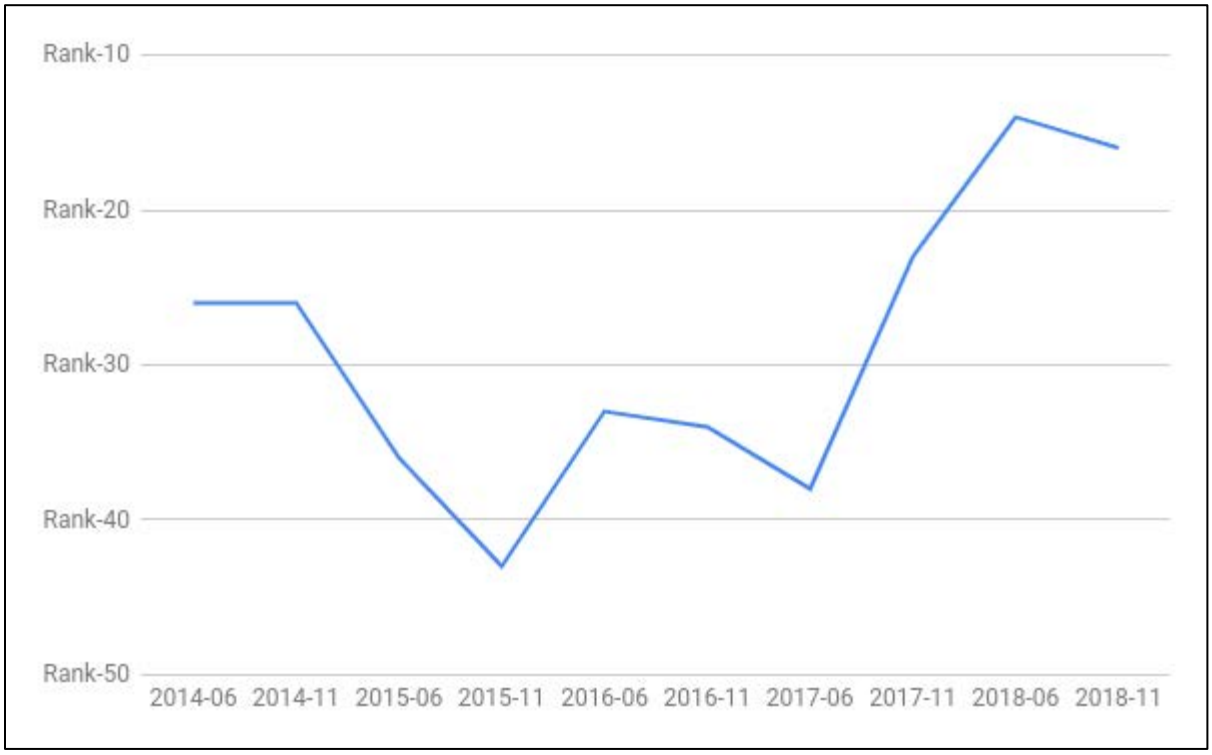

Figure 14: The best rank achieved by a system from Bull in Top500.

2.1.5 Computing efficiency

Figure 15 shows the computational efficiency of the systems in the top 50 positions in the latest released list. The efficiency is represented by the ratio of floating-point operations executed in two popular benchmarks, compared to the theoretical limit (peak performance) of the corresponding hardware. (Please note that some data are missing for the HPCG benchmark.) The graph shows a clear distinction between the values for HPL and HPCG benchmarks, but scores are moderately correlated. In the HPCG benchmark, only one computer in the top 50: The K computer, has achieved a score above 5%, the rest were all below 2.5%: two around 2.4%, five below 1.0%, nineteen provided no data.

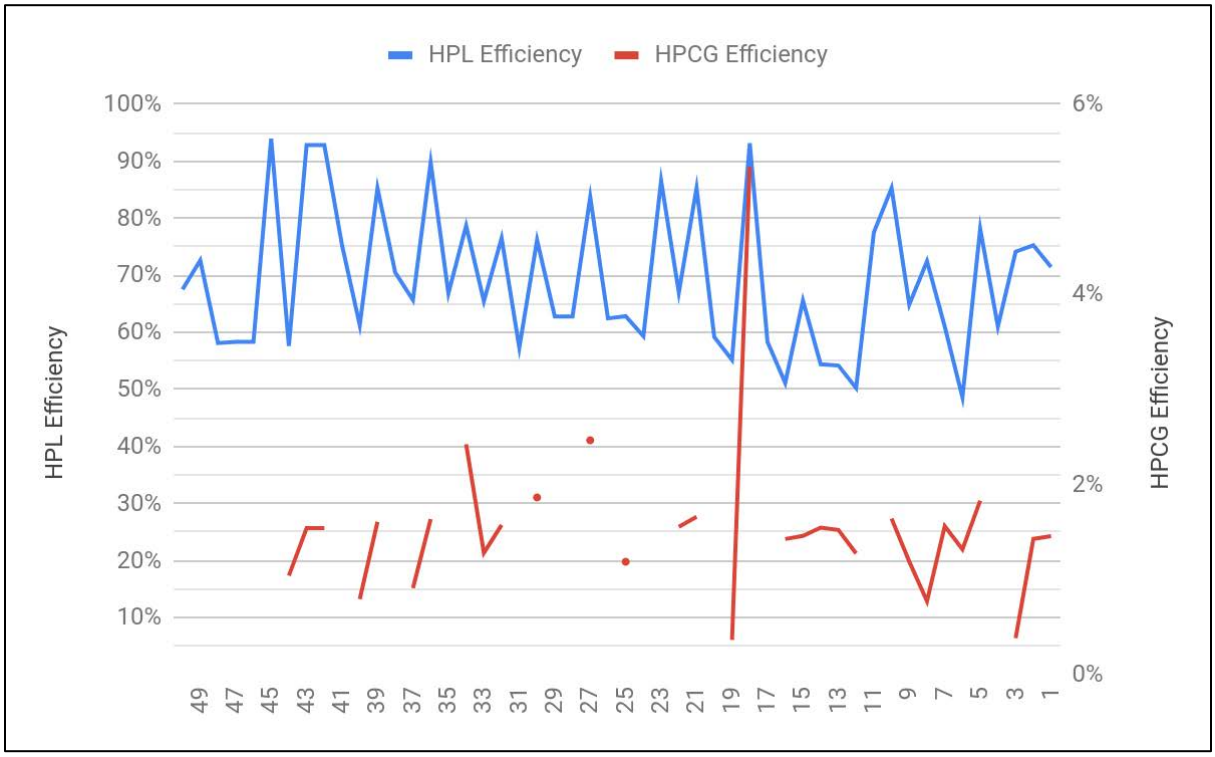

Figure 15: HPL vs. HPCG efficiency in the top50 comparison (% of R_{peak}).

Figure 16 and Figure 17 list the same data for only the top 10 computers, with the additional information of their microarchitectures. For the HPL benchmark, all architectures in the top 10 positions have similar results, with the exception of the Intel Xeon Phi based Trinity having a slightly lower score. The results for the HPCG benchmark are similarly low in variation, with the exception of the ShenWei based TaihuLight, and the Intel Skylake based SuperMUC-NG performing slightly lower than the rest (data for the older generation Intel Xeon based Tianhe-2A is missing).

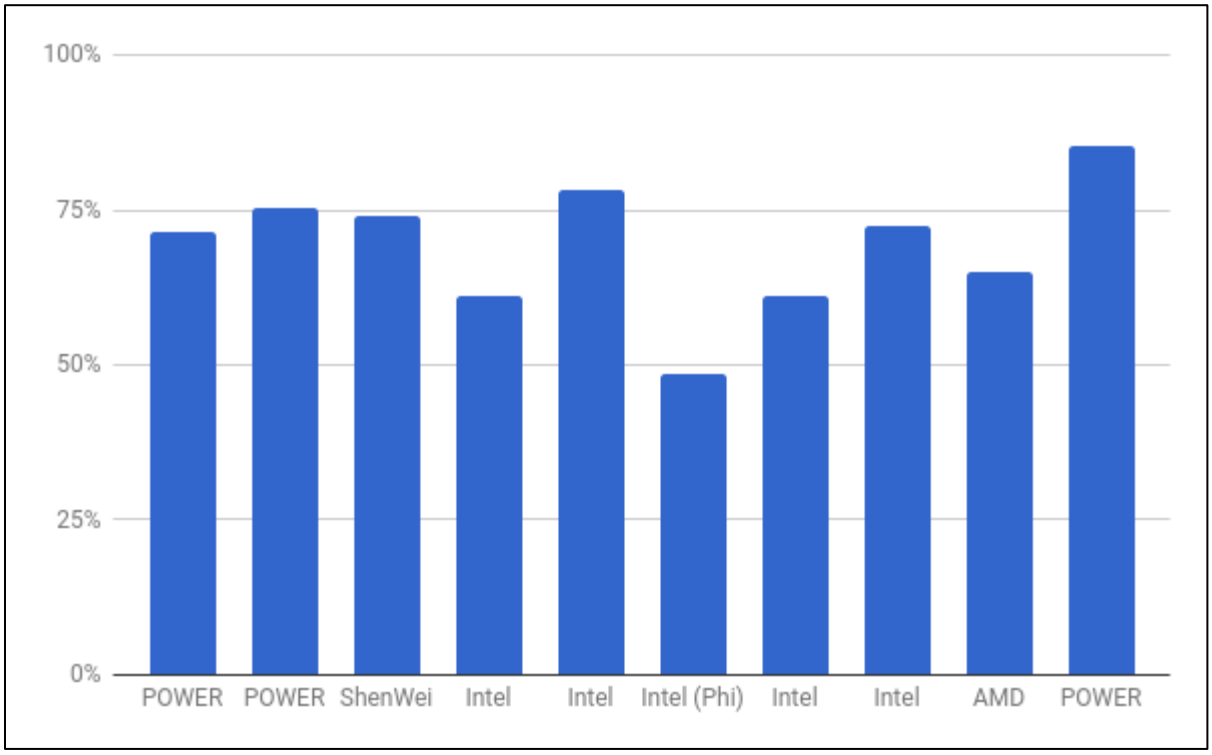

Figure 16: HPL efficiency in Top10 by architecture (% of R_{peak}).

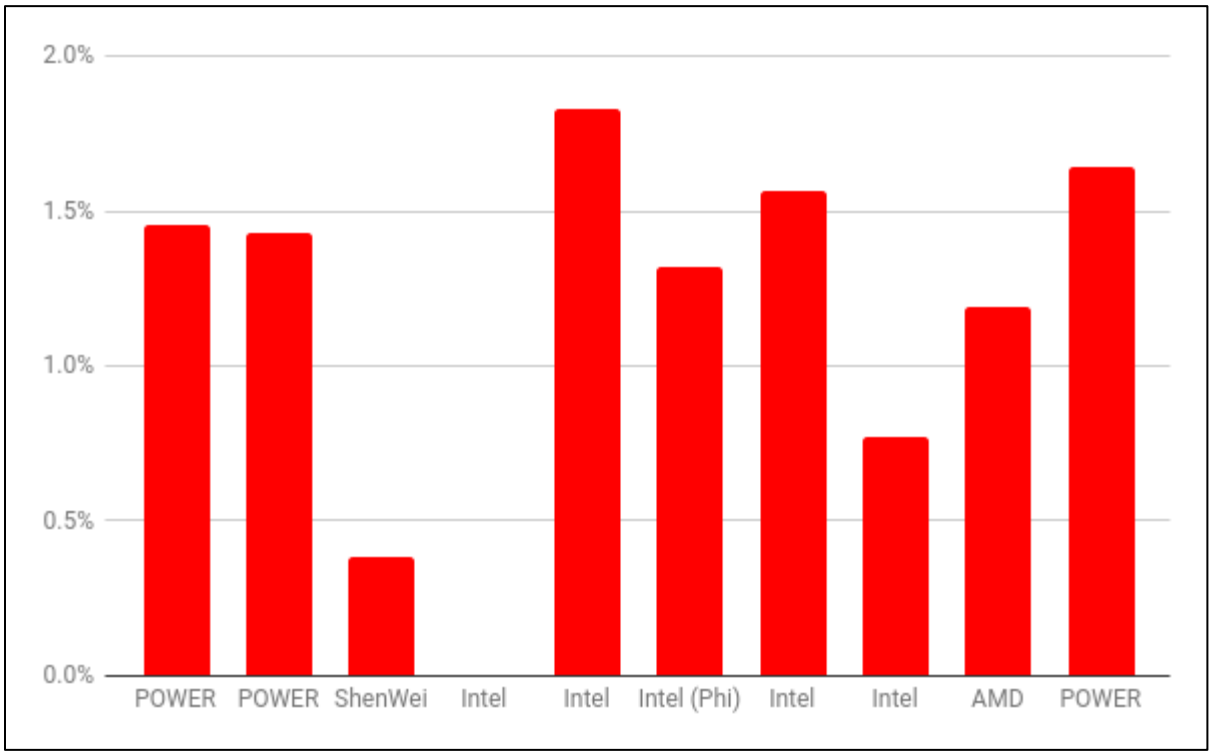

Figure 17: HPCG efficiency in Top10 by architecture (% of R_{peak}).

2.1.6 Energy efficiency

The energy efficiencies of the Top10 and Top50 systems and their Green10/Green50 counterparts are shown in Figure 18 and Figure 19, respectively. In these graphs, GFlop/s/W is used to measure the energy efficiencies. Both figures indicate that there is an apparent increase in the energy efficiency year by year: not only in the Green10, but also in the Top10. The increase in the Green10 starts to become sharper at the end of 2016 and reaches from 5 to 14 GFlop/s/W in 11/2017 and then it flattens until 11/2018. In the same period, the Top10 steadily increases from 3 to 7.5 GFlop/s/W, finally catching up to around %50 of the Green10 at the end of 2018, the same ratio that is observed before Green10's sudden jump during 2016-2017. With 14 GFlop/s/W, the projection to 1 EFlop/s is around 70 megawatts, much better than the expected requirement two or three years ago (around 170 and 220 megawatts, respectively). A denser manufacturing process (such as 10 or even 7 nm) and more efficient accelerators would further lower the energy requirements, making exascale systems more achievable in the future.

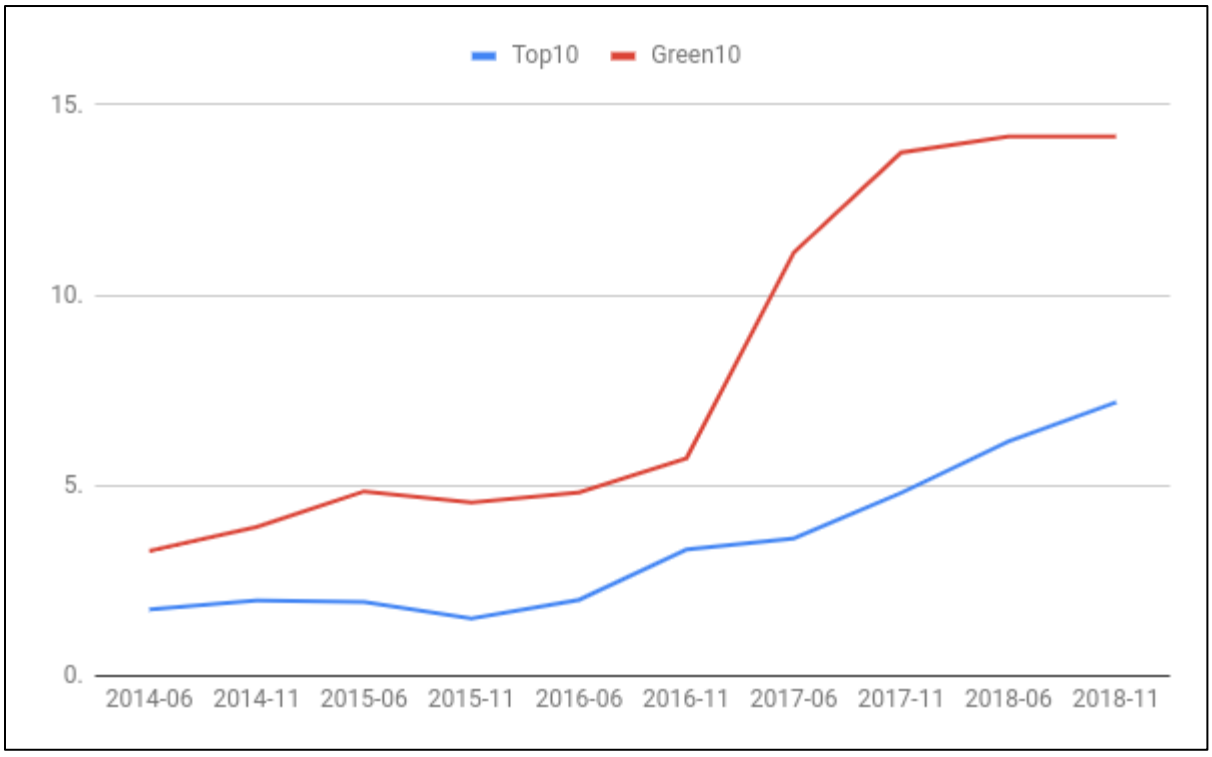

Figure 18: Average energy efficiency [GFlop/s/W] in Top10 and Green10.

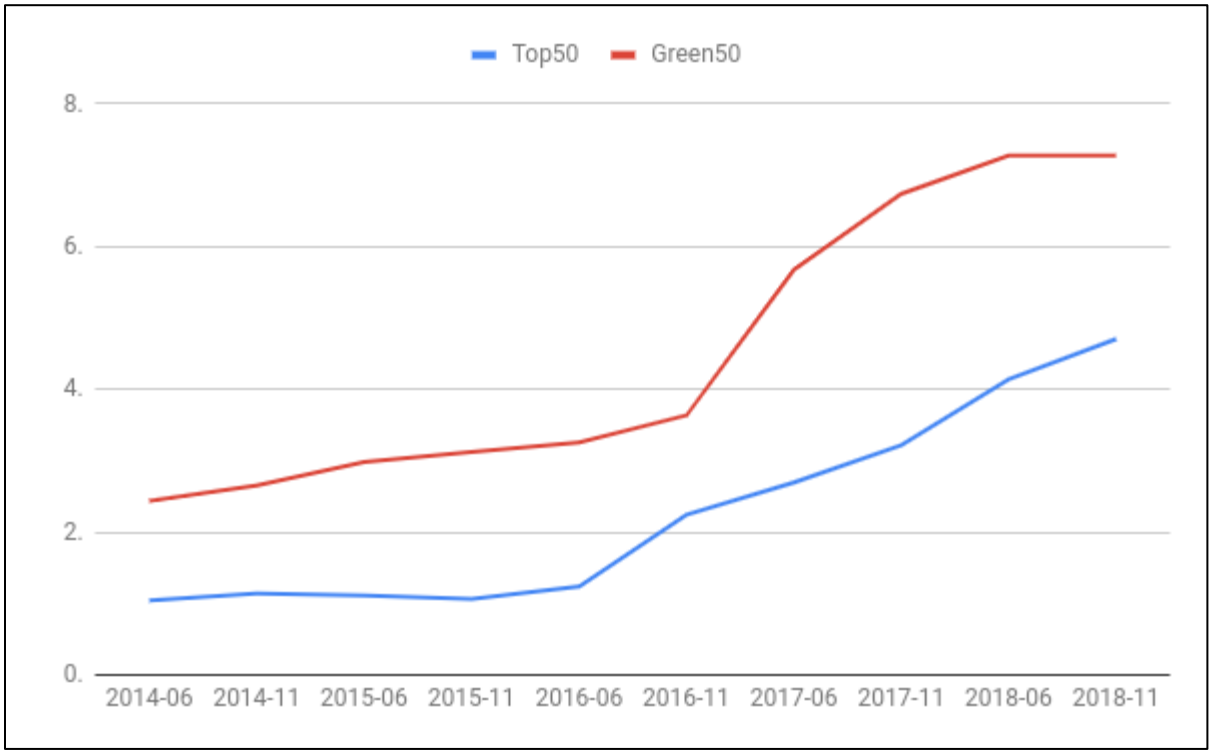

Figure 19: Average energy efficiency [GFlop/s/W] in Top50 and Green50.

2.1.7 Architectures

Figure 20 shows the computer (sub)architectures in the Top10. While AMD x86_64, Intel x86_64 and Intel Xeon Phi share the same general instruction set, they have subtle differences in their extensions (most importantly in vector instructions) and some other design choices and are therefore treated separately here. Intel x86_64 is the most preferred architecture during the last five years, followed by POWER. The sub-architecture AMD x86_64 is represented in the Top10 list by only one machine, TITAN, over the past few years. SPARC disappears from Top500 in 2018. On the other hand, ShenWei entered the list at the end of 2016. The other newcomer is Intel Xeon Phi Knights Landing (KNL), as the main processor instead of an accelerator, since 2016. At the end of 2017, three HPC systems were equipped by Intel Xeon Phi KNL processors and this number is reduced to one at the end of 2018. Figure 21 indicates the best rank by the architectures. Following the fast acceleration started at the end of 2017, POWER + NVIDIA gets the rank 1 in 2018. Before POWER, ShenWei was at the first rank in between 2016 and 2017: today its rank is 3. Intel held the rank 1 only in 2014 and 2015. The rankings of AMD x86_64 and SPARC are in a continuous decline and today they are rank 9 and 18, respectively. Intel Xeon Phi, which appeared first in 2016 at the rank 5, is at the rank 9 in the mid of 2018 and then moved to the sixth position.

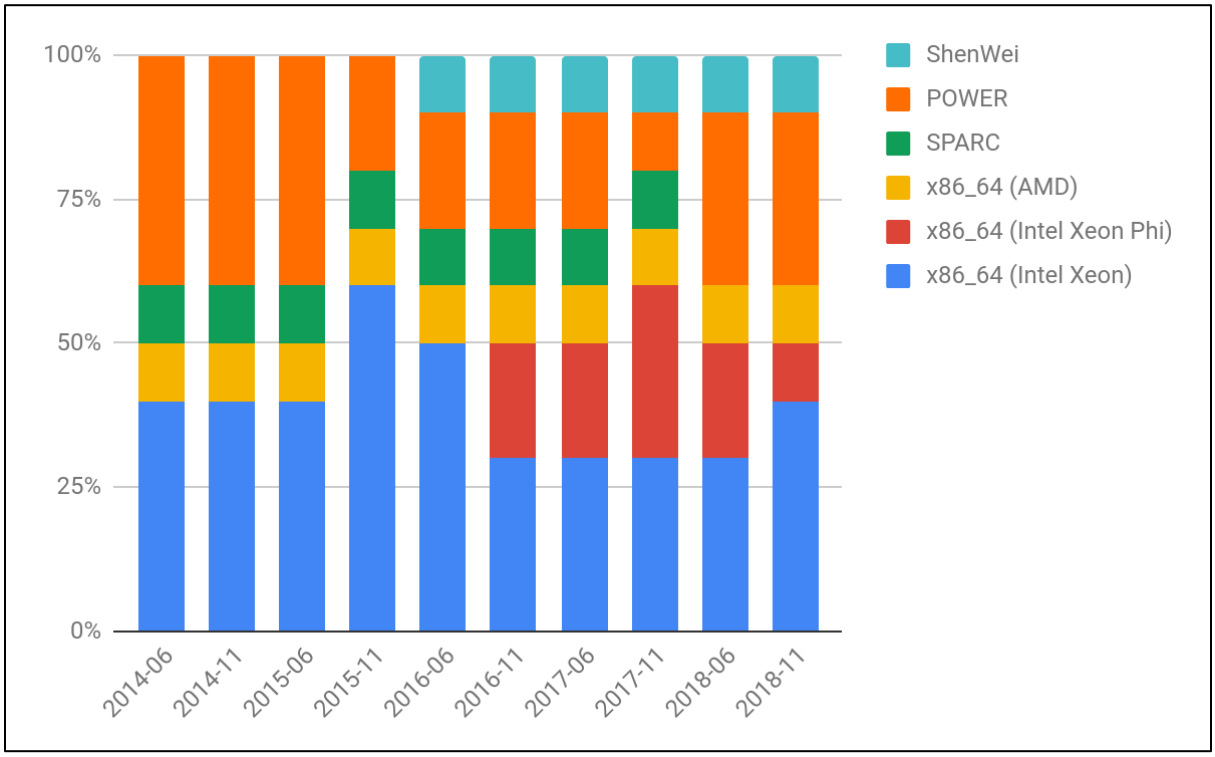

Figure 20: Architectures in the top10.

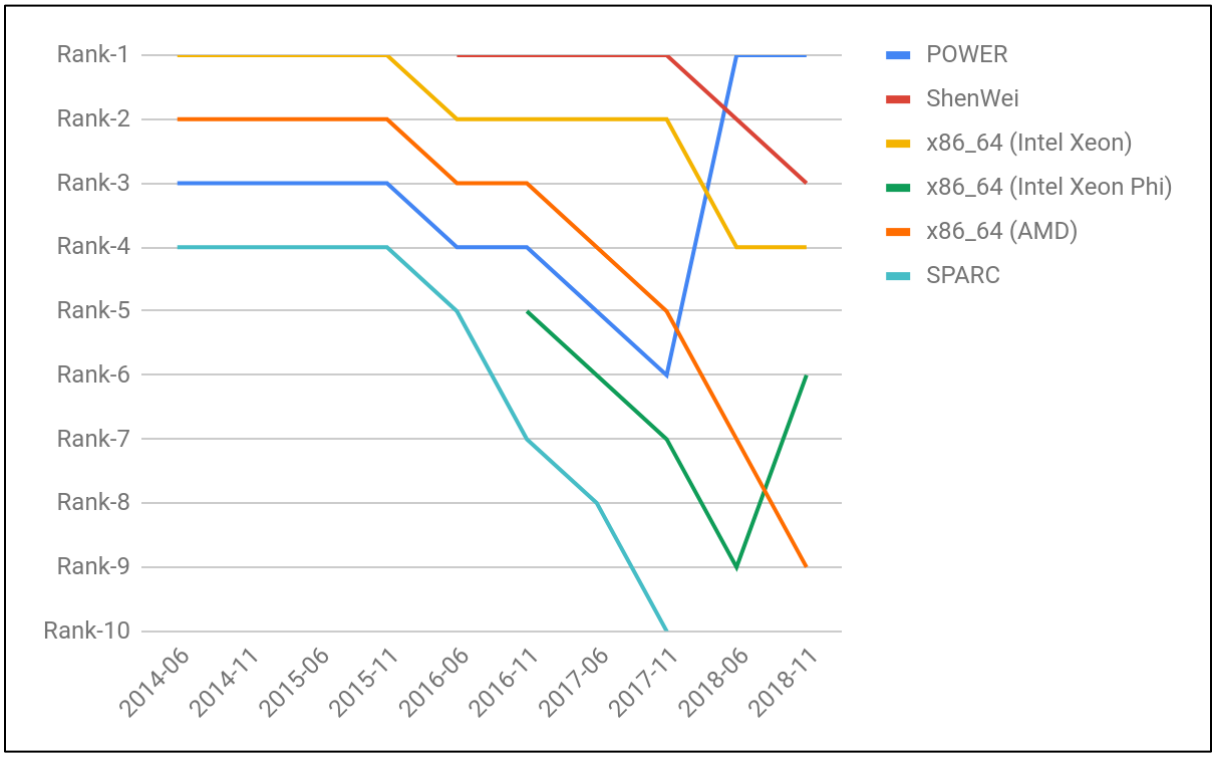

Figure 21: Best ranks by architecture.

2.2 Update on Exascale initiatives

In this short overview, we follow up on the updates of the high-level plans and the underlying technologies with regards to major Exascale plans in Europe and other regions (see previous PRACE deliverables [5][6][7]).

Outside of Europe, three ecosystems dominate the international HPC scene: China, USA and Japan. As a matter of fact, Exascale plans (including financial roadmaps) in these top 4 ecosystems, are clearly dominating the Top500 with more than 90% performance share respectively of the installed HPC platforms. The trend to create synergies between HPC programmes and Big Data and Artificial Intelligence initiatives is confirmed.

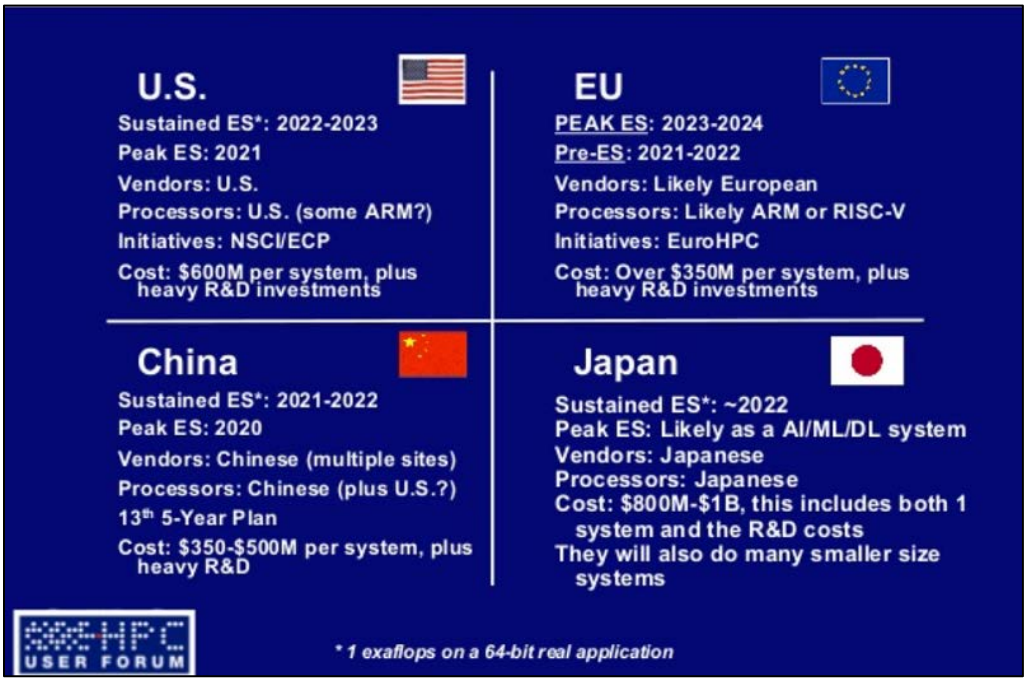

Figure 22: Exascale initiatives schedule and efforts from US, EU, China and Japan (source Hyperion)

2.2.1 Exascale plans: China

After a long leadership era over Top500 #1 ranking (from June 2013 to November 2017), China "only" holds the first rank in system shares (227/109) in November 2018 list, becoming also 2nd for performance share (31% / 37.7%) behind the USA. China now masters HPC processor and interconnect technology and confirms its plans to achieve exascale through the announcements at SC18 in Dallas.

Depei Qian, chief scientist of China's national R&D project on high performance computing, presented at SC18 an overview of the country's exascale plans, relying on China controlled technologies (i.e. most if not all of the hardware and software elements developed domestically).

Three prototypes – Sugon, Tianhe, and Sunway (ShenWei) – were deployed over the 10 first months of 2018, with the last one being unveiled just a before SC18. Neither of them appears to be based on x86 technology.

Outline of goals:

- peak performance of one peak EFlop/s with HPL efficiency > 60%;
- minimum system memory capacity of 10 PB;
- interconnect at HPC grade latency with >100GB/s node-to-node bandwidth;
- energy efficiency of at least better than 30 GFlop/s per Watt. (close to 20 – 30MW being envisioned in exascale programs in the US, Japan, and the EU);
- Large scale system management and resource, system monitoring and fault tolerance
- Exascale targets based on these prototypes are (still speculative):
  - 2020 for Tianhe-3 (in Tianjin), announced to be first;
  - 2020 – 2021 for Sunway and Sugon (Shandong area).

Co-design of such 3 architectures is based on the key following application domains:

- Numerical nuclear reactor
- Numerical aircraft (CFD, structure, MDO) and engines (4-stage unsteady LES simulation)
- Astrophysics and earth system
- Drug discovery and life sciences
- Seismic and oil exploration
- Material sciences
- Complex engineering (ex: 3 Gorges full dam modeling)

2.2.1.1 SUGON

Based on AMD-licensed Hygon x86 processors, enabling legacy code support, associated with an accelerator "DCU", targeting a node (2 procs + 2 DCU) @15 TFlops. Interconnect is a 6D Torus, based on 200Gbps technology, with a 500Gbps exascale goal. The Sugon machine uses an immersive cooling system, boiling at the low temperature of 50°C.

2.2.1.2 TIANHE-3

The system should be based on a Chinese-designed ARM chip, likely be some version of Phytium's Xiaomi platform within a reconfigurable flexible heterogeneous architecture. As with the Sugon prototype, the Tianhe system is made up of 512 nodes and delivers nearly identical performance of 3.14 PFlop/s. That suggests quite a powerful processor. The interconnect is a 3D butterfly design with fault tolerance as a key design feature. It is expected that Tianhe-3 will be in 2020 close to 200 times faster and have 100 times more storage capacity than the Tianhe-1 supercomputer.

In July 2018, the Tianhe-3 prototype has completed acceptance testing for China's Ministry of Science and Technology, followed by the acceptance of a second prototype from Sunway.

2.2.1.3 SUNWAY (ShenWei)

This prototype relies on ShenWei 26010 (SW26010) 260-core processor, that powers the Top500 #3 (November 2018) Sunway TaihuLight. Each node is based on two of these processors, that

together deliver about 6 peak TFlop/s. The 512-node (dual socket) machine reaches 3.13 PFlop/s peak performance.

Energy efficiency is around 11 GFlop/s per watt. Tripling is expected to meet the stated target for exascale in three years, which will require a major innovation.

The prototype employs a home-grown network chip providing 200Gbps of point-to-point bandwidth (as Sunway TaihuLight uses Mellanox).

2.2.2 Exascale plans: Japan

In Japan, the most important effort remains the post-K project, targeting a massively parallel exascale-class supercomputer in 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 work in nine application domains.

Fujitsu has announced in August 2018 the specifications for the ARM CPU A64FX: 8,786 million transistors at 7nm process technology. It will be the first CPU to implement ARM's Scalable Vector Extension (SVE), an instruction set designed specifically for high performance computing. Fujitsu has already produced a prototype of the processor and has started its initial testing. The core count is 48 compute cores plus 4 assistant cores associated with the SIMD vector 512 bits width.

The A64FX delivers 2.7 TFlop/s FP64, over 5.4 Flops FP32, and over 10.8 Flops for FP16. The latter two are especially important for deep learning applications for training neural networks.

The A64FX also has implemented integer dot product operations for 16-bit (INT16) and 8-bit (INT8) formats, which can be used for inferencing in neural networks. Fujitsu announced that the new CPU can achieve more than 21.6 TOp/s using INT8 and more than 10.8 Tops for INT16.

The A64FX will be equipped with an on-chip network controller that will route data over the system's Tofu fabric. For Post-K, this fabric will be a 6D mesh/torus with each processor providing 2 lanes, each with 10 ports at 28 Gbps. That works out to 560 Gbps per CPU or node.

The A64FX will deliver up to 1024 GB/second per CPU using 32GB of on-package HBM2 memory. According to Fujitsu, they are able to achieve over 830 GB/second on the Stream Triad benchmark, a yield of over 80 percent of the processor's peak bandwidth.

Post-K will provide up to 100× performance of K on real applications with a converged and balanced architecture targeting HPC + AI applications (fp16, int8).

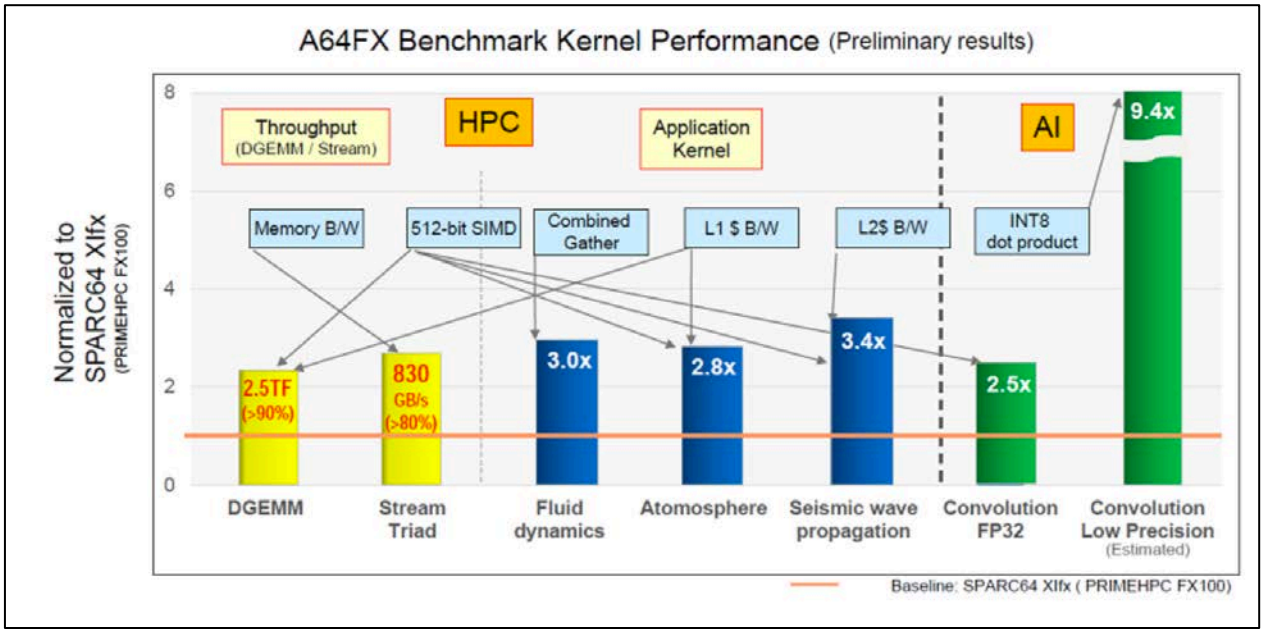

Figure 23: A64FX Benchmark Kernel Performance preliminary results

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, Juelich, RIKEN). The software stack developed at RIKEN is open source. It also runs on the Intel Xeon architecture. By 2021, a full stack of HPC software components is expected to be available, including Linux, C/C++ and Fortran compilers, debuggers, MPI, OpenMP, math libraries, resource managers, and Lustre.

Like all previous Japanese HPC systems since decades, Post-K is developed by co design using 9 applications: GENESIS (md for proteins), Genomon (geonome processing), Gamera (earthquake simulator), NICAM+LETK (weather prediction using big data), NTChem (molecular structure calculation), FFB (LES), RSDFT (ab initio DFT), Adventure (computational structure mechanics) and CSC-QCD (Lattice QCD).

As mentioned in D5.1, Japan has launched an ambitious plan in AI (more than $1B) and the HPC community works on how to leverage HPC technologies for this field [5].

2.2.3 Exascale plans: USA

US Exascale milestones are:

- 2021 (A21) at Argonne National Lab 1 EFlop/s peak ~80% Linpack ratio, 40 GFlop/s/W, provided by Intel and Cray,
- 2023 (Frontier, Summit follow-on) at Oak Ridge National Lab, range 1.5 to 3 EFlop/s peak ~70% Linpack ratio, 60-100 GFlop/s/W,
- 2024 (El Capitan, Sierra follow-on) at Lawrence Livermore National Lab, range from 4 to 5 EFlop/s ~60% Linpack ratio 130-200 GFlop/s/W.

Five years after the spectacular first #1 Top500 of China in June 2013, USA won back the “summit” of Top500. In June 2018, SUMMIT took the lead and in November 2018, SUMMIT was slightly

upgraded (143.5 PFlop/s R_{max}) and SIERRA (94.6 PFlop/s R_{max}) entered rank 2. Some numbers about SUMMIT, which is likely to draft the profile of an exascale configuration:

- Two tennis courts in surface, 9.8 MW, 340 tons, direct liquid cooling at 21°C
- Green500 energy efficiency is around 15 GFlop/s/W.

Respectively run at Oak Ridge National Lab [9] and Lawrence Livermore National Lab [10], SUMMIT and SIERRA, both machines are OpenPOWER technology based (IBM POWER9 CPUs tightly coupled with NVIDIA Volta GPUs). They are winning steps in President Obama's National Strategic Computing Initiative (NSCI) [11] in 2015 aiming at giving US HPC high political visibility and regain scientific simulation supremacy. Despite the fact that the US lost the Top500 leadership in terms of the number of systems, they still dominate the Top500 as far as US technology in HPC systems is concerned.

Within NSCI, the program lead by DOE is the Exascale Computing Initiative (ECI) including the main efforts for the delivery of the exascale computing capability organised under the Exascale Computing Project (ECP) [12]. As already said in D5.1, 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). ECP covers the development of both hardware and software technologies, systems and applications - procurements of exascale systems will follow SC and NNSA processes and timelines.

ECP has four focus areas:

- Application development
- Software technology
- Hardware technology
- Exascale systems testbeds

Main objectives are:

- Parallelism: 1000-fold larger than Petascale systems
- Memory and storage: provide access in line with the anticipated computational rate
- Reliability: resiliency at system and application level
- Energy efficiency: to stay below 20-40 MW for an exascale system.

The Exascale Computing Project has initiated its sixth Co-Design Center, ExaLearn (announced in June 2018), to provide exascale ML software for use by ECP Applications projects, other ECP Co-Design Centers, DOE experimental facilities, and leadership class computing facilities. The *ExaLearn* Co-Design Center will also collaborate with ECP PathForward vendors on the development of exascale ML software.

2.2.4 Exascale plans: Europe

Europe has an ambitious plan to become a main player in supercomputing (motivated by strategic considerations and the discrepancy between the European consumption of HPC resources worldwide today (29%), and the EU industry providing only ~5% of HPC platforms). The EuroHPC initiative is a joint undertaking with as one of its goals to construct an exascale supercomputer based on European technology [13] and fund a world-class European supercomputing infrastructure during 2018–2026 to meet the demands of European research and

industry. The first ambitious objective of EuroHPC is to acquire at least two pre-exascale computers by 2020/21 and to reach full exascale performance by 2023. The objective is to define testbeds for HPC and big data applications that will make use of these supercomputers for scientific, public administration and industrial purposes.

EuroHPC strategy includes processor design and production. A significant effort has started to build a production HPC processor with industry grade: "European Processor Initiative" EPI. This is done as part of a 120-million-euro Framework Partnership Agreement (FPA). The EPI consortium consists of 23 partners from industry, research, and education, among which: Atos Technologies, E4 Computer Engineering, Extoll, Infineon, Kalray, and ST Microelectronics among the IT suppliers; Barcelona Supercomputing Center, CEA, Cineca, GENCI, Research Centre Julich, ETH Zurich, Forth Institute of Computer Science, Technico Lisboa, and Universita di Pisa from the national supercomputing labs and academic labs; and BMW Group and Rolls Royce from the commercial HPC users, among others.

For the European Union, Exascale in not only putting together a system that can hit an EFlop/s on Linpack, but it is all about building a system where the exascale capacity is usable with real world application performance, and the building blocks and components should address mass market and industry needs.

The EPI roadmap indicates that their current plan is to couple ARM CPUs with RISC-V (pronounced *risk five*) accelerators to produce a first generation pre-exascale system by 2021. Tape out of the first generation processors should take place in 2019.

The second generation exascale system will come out in the middle to late 2023. The timing is approximate and a bit hazy on the chart. The first automotive proof of concept will debut in 2022 and be commercially available in the middle of 2024.

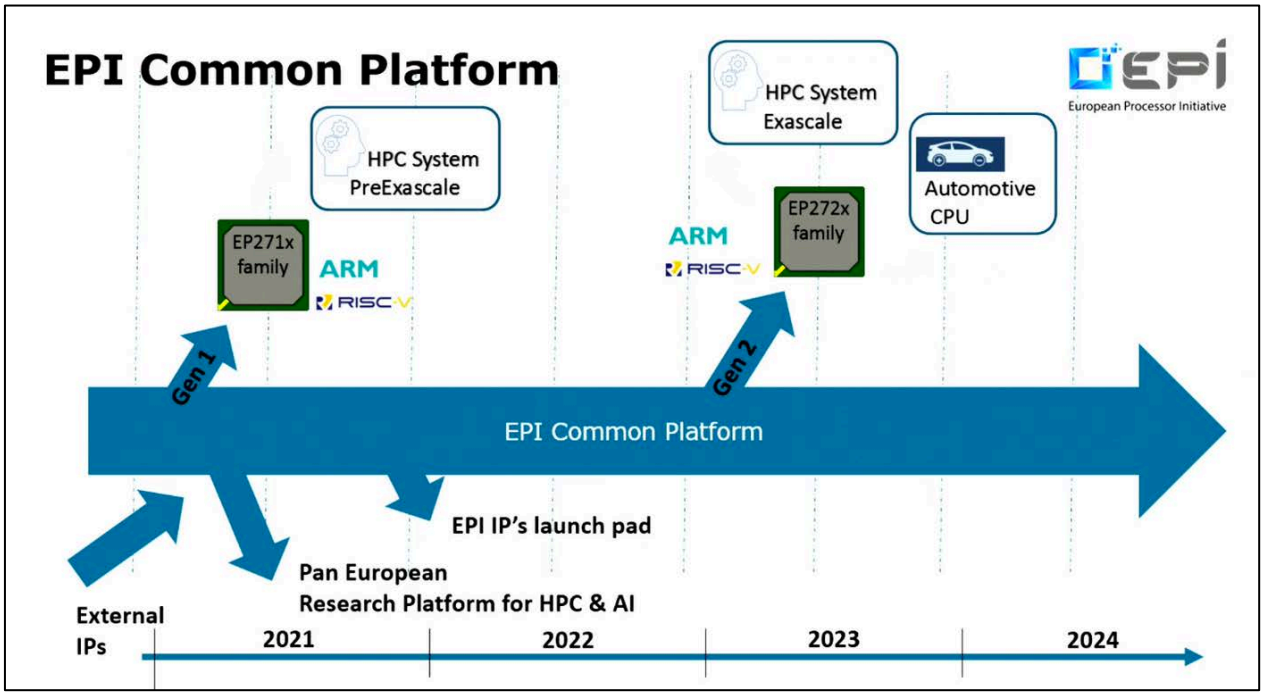

Figure 24: European Processor Initiative roadmap

2.3 Business analysis

According to Intersect360 Research [14], the total HPC market revenue was $35.4 billion in 2017, up by 1.6% from 2016. Cloud grew by 44% from 2016, reaching $1.1 billion revenue in 2017. All categories that comprise Cloud (Raw cycles, Storage, Application Hosting-SaaS, Infrastructure Hosting-IaaS/PaaS, Other) grew in 2017, with SaaS growing by a staggering 125%. Storage and servers grew 3% and 7%, respectively. Networks, Software and Services declined in 2017 (Figure 25).

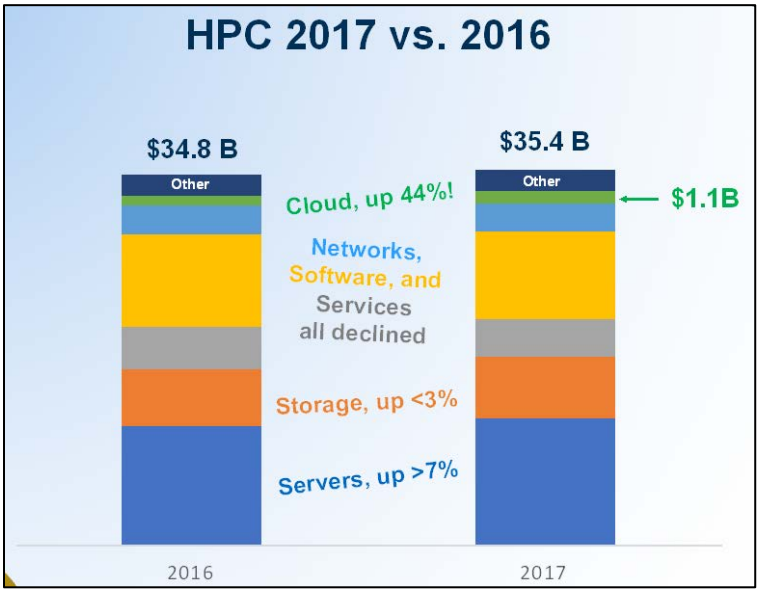

Figure 25: HPC Market growth (2017 vs 2016) from Intersect360 Research

In 2017, roughly $4.5 billion were spent for Machine Learning (ML), 90% of which came from hyperscalers. The most advanced use of ML cases within HPC come from the finance sector. Moreover, 56% of the HPC organizations are utilizing Machine Learning, usually on the same hardware as HPC, or on GPUs.

Table 3 shows the HPC 2017 revenue by vertical sectors. The 56.5% of the revenue comes from the Commercial Sector (e.g. large product manufacturing, bio-sciences, energy, consumer product manufacturing, etc.), while Government sector (national security, national research labs, etc.) follows with 25.7% and Academic/not-for-profit is last with 17.7%.

Vertical Sector	% of 2017 Revenue
Commercial	56.6%
Government	25.7%
Academic	17.7%

Table 3: HPC 2017 Revenue by Vertical (Intersect360)

For the future, Intersect360 predicts that the Cloud will continue its fast growth in HPC, reaching approx. $2 billion in 2019 and approx. $3 billion in 2022. Furthermore, almost all HPC Cloud usage will be hybrid, combining on-premise HPC capabilities with remote HPC in the cloud.

According to Hyperion Research [15], the Worldwide HPC Server Market for 2017 was $12.3 billion (Figure 26), which comprises of $4.6 billion revenues from Supercomputer (over $500K),

$2.3 billion from Divisional HPC ($250K-to-500K), $3.5 billion from Departmental HPC ($100K-to-250K) and $1.9 billion from Workgroup HPC (under $100K) (see Table 4).

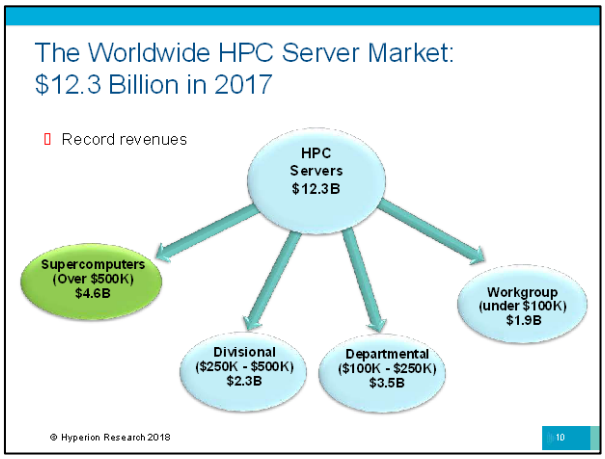

Figure 26: Worldwide HPC Server Market from Hyperion Research.

OEM	2017 (k$)
HPE/HP	4,194,470
Dell	2,330,134
Lenovo	869,895
Inspur	740,207
IBM	575,130
Sugon (Dawning)	348,846
Cray	250,195
Fujitsu	227,802
NEC	171,344
Bull Atos	133,422
Other	2,420,852
Grand Total	**12,262,296**

Table 4: HPC Server Marker by vendor from Hyperion Research

The Worldwide HPC Server Market is expected to grow by 9.8% (Compound Annual Growth Rate) from 2017 to 2022. In Table 5, the Total HPC Server Market growth is shown, as well as the individual subdivisions.

	2017 (k$)	2022 (k$)	CAGR 2017-2022
Supercomputer	4,840,985	9,521,092	14.50%
Divisional	2,295,273	3,058,506	5.90%
Departmental	3,370,137	4,540,846	6.10%
Workgroup	1,755,901	2,436,173	6.80%
Total	**12,262,296**	**19,556,617**	**9.80%**

Table 5: Worldwide HPC Server Market Forecasts by Hyperion Research

According to Hyperion Research, the broader HPC Market will grow by 9.6% (CAGR) from 2017 to 2022, with the respective subdivisions growing from 6.7% to 10.30%, as shown in Table 6.

	2017 (k$)	2022 (k$)	CAGR 2017-2022
Server	12,262,296	19,556,617	9.80%
Storage	4,661,102	7,606,884	10.30%
Middleware	1,367,147	2,170,912	9.70%
Applications	3,961,491	6,227,996	9.60%
Service	2,022,090	2,798,378	6.70%
Total Revenue	**24,274,126**	**38,410,788**	**9.60%**

Table 6: Broader HPC Market Forecasts by Hyperion Research

Hyperion Research predicts that 2022 will be a very strong year for HPC market, while 2018 did very well with growth that might have exceed 8%. Cloud computing, HPDA, including ML/DL, cognitive and AI are fuelling the growth, driving storage growth and new types of technologies and expanding opportunities for vendors.

All the data from Intersect360 Research and Hyperion Research were gathered during public presentations, briefings and events that took place during the 2018 ISC High Performance (Jun 24–28, 2018, Frankfurt, Germany) and 2018 SC Conference (Nov 11–16, 2018, Dallas, Texas).

2.4 Cloud computing in HPC

HPC targeted cloud services have already been available for some time. Big players like Amazon, Microsoft and Google have offerings that target the HPC market. OpenStack is also popular within research organizations and several traditional HPC centers already offer a solution on top of OpenStack. These generally complement, rather than replace the traditional HPC systems and provide a computing environment to the long tail of applications.

Using commercial cloud providers is a real option for some HPC workloads. In small to medium-sized use cases, where there is no long-term need for resources or where the need varies heavily, commercial cloud providers start to have solutions. In addition to traditional IaaS offering, cloud providers also have PaaS and SaaS solutions where customers can provision a complete cluster with software for a particular workflow. Major cloud providers also have market places where smaller HPC software companies are providing their solutions tailored to run in the cloud. These infrastructures will still struggle to support high-end use cases, which are normally run on Top 500 machines. In addition, for long term use, the cost of solutions based on commercial cloud is likely higher than using a dedicated system.

OpenStack has been the largest and most deployed in-house cloud platform for a while. Generally, OpenStack has the required functionality to build an HPC cloud, and it has been used in several cases. Using technologies like PCI pass-through, Single Root I/O Virtualization (SR-IOV) and bare-metal provisioning one can build an HPC service with the desired balance of flexibility and performance. There are few if any technical limitations for building an in-house HPC cloud any more. The problem has moved towards a business question of cost/benefit analysis of an 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 an HPC cloud service is higher than a traditional HPC service, but it adds flexibility.

Traditionally, HPC cloud discussions have revolved around IaaS. Lately, this has been changing. The container orchestration engine Kubernetes [16] has developed quickly, and offers cloud services on a higher abstraction layer. Its use of containers allows for easier mitigation of some cloud performance impacts compared to virtualized clouds and also helps portability of applications between HPC systems. 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. Use of container technology for portable workloads is currently replacing grid computing and grid middleware for distributed computing. For workloads using traditional batch scheduling, tools like Singularity [17] and Shifter [18] provide the benefits of the containerization, although they should be considered more like application deployment on traditional HPC systems rather than cloud services.

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.

2.4.1 Trends

2.4.1.1 Bare metal services

Bare-metal cloud services seem to be a trend in 2018. AWS, Oracle, Alibaba and Rackspace released bare metal HPC services, and HPE partnered with Advania to provide HPC cloud services. The latter does, however, struggle a bit with the cloud definition, as it seems there is no self-service option, but you have to go through a sales process.

Nonetheless, it seems that cloud providers see value in bare metal HPC services. In order to utilize the full performance of NVMe based, low latency local storage requires bare iron access to hardware and using high bandwidth networking e.g. 100Gb InfiniBand, if provided, requires utilization of full compute nodes and running virtual machines on top of the hypervisors gets irrelevant. But there is also a downside with bare metal access – many of the benefits of virtualization like redundancy and enhanced information security are lost and because of low-level access from users, some storage and network products are not available in the bare metal. The actual benefit of bare metal should become more visible as you scale, which will make it interesting to monitor if the providers can provide the services at a scale and a price point where it generates enough interest.

2.4.1.2 GPUs and FPGAs

It seems like most cloud providers were eager to jump on the GPUs for AI acceleration bandwagon. In addition to the large providers releasing instances equipped with NVIDIA V100 cards, so did Oracle and IBM. It seems like GPUs are quite a standard offering for most cloud providers, and new instance types will be released by most vendors as the technology becomes available.

The availability of FPGAs in the cloud seemed to make a larger splash a few years ago. While the availability of FPGAs grew in 2018, on the whole, the large growth of GPUs easily outpaced the availability and use of FPGAs. The technology still offers several benefits, so if the supported application list grows, they might soon become interesting to a larger audience.

2.4.1.3 Quantum computing

Quantum computing is an ideal use case for cloud since actual quantum devices are still experimental or prototypes, very expensive and often require special hosting infrastructure. The wall time of quantum computations is very short, so the resource starvation is not a problem. A cloud interface makes it possible to make physical resources available to a wide range of research and business. Typically, the cloud interface includes a software development kit and job scheduling function.

IBM has 5- and 14-qubit Q systems available in the cloud for public use and a 20-qubit Q system for IBM customers. Rigetti Computing is providing a full stack quantum computing cloud service including a software development kit and traditional computing resources tightly integrated to a 16-qubit quantum processing unit Aspen-1. CAS-Alibaba Quantum Company Laboratory provides access to an 11-qubit superconducting quantum computing device. D-Wave Systems launched in October 2018 Leap, a cloud service providing real-time access to a D-Wave 2000Q quantum annealing system including Ocean SDK. Google has stated to have their upcoming, 72-qubit "Bristlecone" QPU available in the cloud in near future.

2.4.1.4 Custom and heterogeneous hardware

In 2018, a significant amount of custom hardware was introduced by the large players. AWS announced their ARM servers, Google released the TPU 3.0 for AI workload acceleration and Facebook reinforced its investments within the Open Compute Architecture with its Zion server (8 way GPU or Intel Nervana). In addition, AWS has their Nitro ASICs [19] as the next step in their management of the hardware. Nitro ASICs and hypervisors should allow offloading more functionality from the hypervisor to a dedicated card, thus improving performance and performance predictability.

The hardware heterogeneity also seemed to grow in 2018. In addition to vendors releasing GPUs, FPGAs and custom hardware, AWS also released AMD Epyc based instances and Nimbix released an offering with POWER9 processors.

2.4.2 European Open Science Cloud

The European Open Science Cloud (EOSC) was launched in late 2018. Its aims are not to offer cloud services by itself, but rather be a place where communities and platforms can meet. Currently, at the time of the launch it has some cloud services available, but so far it is taking its first steps. As EOSC matures, it might attract more and more services. EOSC is not aimed to be a purely HPC centered solution, but due to its central position, in the long run, EOSC may turn out to be a relatively important hub for connecting to scientific cloud services, including a selection of HPC offerings.

2.5 Consolidation in the HPC market

The last year has brought several acquisitions and mergers in the market that may improve the quality of products for HPC. The vendor market is reshaping due to current IT trends. In this section, the main consolidations in the hardware (server/storage and semiconductor) and software areas are outlined:

- Marvell acquisition of Cavium – Cavium's portfolio of multi-core processing, networking communications, storage connectivity and security solution extend Marvell's storage, networking solutions and high-performance wireless connectivity products [20].
- Broadcom acquisition of CA Technologies – CA Technologies IT management software and solutions extends Broadcom portfolio [21].
- Arista Network acquisition of Metamako – Arista cloud networking solutions for large datacenter and campus environments, has acquired Metamako provider of low-latency, FPGA-enabled network solutions [22].
- HPE acquisition of PLEXXI Software Defined Network Data Fabric [23].
- HPE acquisition of CAPE Sensor Based Network Performance and Monitoring [24].
- Wave Computing acquisition of MIPS – Wave Computing specializing in data flow processing of Deep Neural Networks has acquired MIPS Technologies provider of RISC processor architectures and IP cores [25].
- One Stop Systems acquisition of Bressner Technology – One Stop Systems (OSS), a company that manufactures specialized high performance compute accelerators, flash storage arrays and customized servers, has acquired Bressner Technology, a standard and customized servers, panel PCs, and PCIe expansion systems provider [26].
- IBM acquisition of Red Hat – Red Hat is a leading provider of enterprise open source software solutions, using a community-powered approach to deliver reliable and high-performing Linux, hybrid cloud, container, and Kubernetes technologies extend the IBM portfolio [27].
- DataDirect Networks acquisition of Tintri Inc. – DDN supplier of high-performance data management solutions has acquired Tintri that offers an enterprise cloud infrastructure and all-flash storage with scale-out and automation [28].
- Xilinx Inc. acquisition of DeePhi Tech – Xilinx adaptive and intelligent computing provider has acquired DeePhi specializing in deep compression, pruning, and system-level optimization for neural networks [29].
- DataDirect Networks acquisition Lustre File System Capability from Intel – DNN has acquired Intel's Lustre File System business and related assets [30].
- Oracle acquisition of Talari Networks – Talari Networks' Software-Defined Wide Area Network (SD-WAN) products extend Oracle's portfolio [31].
- Oracle acquisition of DataFox – Datafox a cloud-based AI data engine enhances its cloud applications [32].
- Oracle acquisition of DataScience.com – DataScience.com solutions extends Oracle's cloud department [33].
- Microsoft acquisition of GitHub – GitHub will operate independently and remain an open platform [34].

3 EU HPC Landscape and Technology Update

3.1 EU landscape overview

This chapter introductory section is an overview of the recent and on-going evolution of the European HPC landscape. EuroHPC is the central and overarching new element, consolidating or recomposing the ecosystem. This is already well documented and periodically updated in PRACE WP2 deliverables in particular (such as those on Stakeholder Management [35]). So, we will remain at a very general level in this section, briefly highlighting the elements, players, initiatives and projects that relate more particular to technology, since this is the focus of this report. We will often refer to external documentation from other projects or entities (such as PRACE, ETP4HPC, EXDCI), rather than repeating them with too many details.

In particular, a complete catalogue of H2020-funded HPC projects can be found in [36]. This publication details the HPC Technology, Co-design and Applications Projects within the European HPC ecosystem, including also the EPI.

Some elements of impact assessment of the EU HPC Programme, with some focus on technologies, can also be found in the HPC cPPP Progress Monitoring Report published Fall of 2018 (covering the 2014-2017 period [37]).

3.1.1 EuroHPC

The EuroHPC Joint Undertaking is now getting underway [38][39]. Its first Governing Board meeting was held on Tuesday 6th November 2018. The main strategic objectives of EuroHPC are:

- to equip the Union with the computing performance needed to maintain its research at a leading edge, meeting the computing and data processing needs of European scientists and industry, by deploying a world-class pre-exascale, then exascale, HPC infrastructure in Europe;
- and to support the development of HPC technologies and applications across a wide range of fields (via an ambitious research and innovation agenda to develop and maintain in the Union a world-class High-Performance Computing ecosystem, exascale and beyond).

The usual breakdown of the HPC value chain and related ecosystem into three pillars – technologies, infrastructure, applications– is still fully valid: EuroHPC is a new step towards a more coordinated vision and more integrated strategy for the whole of these three areas. The mapping of the three areas onto EuroHPC pillars is not one-to-one: infrastructure-related aspects (acquisition and operations of supercomputers) make up the first pillar of EuroHPC, whereas technologies and applications are dealt with in the R&I pillar of EuroHPC - whose full scope is: exascale technologies and systems (incl. low-power processor); applications; and skills.

3.1.2 PRACE

PRACE welcomes the creation of the EuroHPC Joint Undertaking and recognises the importance of an appropriate relationship at the scientific policy level and an appropriate coordination between the evolution of PRACE and the creation and development of the EuroHPC JU. In light of this, PRACE published a White Paper on the role of “PRACE in the coming EuroHPC Era” which

describes a proposal for this coordination [41]. PRACE has always had a strong link with technologies (technology watch, deployment of prototypes, pre-commercial procurement) and applications, supporting and enabling them on production but also experimental systems – contributing to filling the gap between emerging technologies and mature technologies, which need to be deployed and harnessed at large production scales by applications.

3.1.3 ETP4HPC, the HPC cPPP, and BDVA

Two other existing stakeholders having activities related to HPC technologies are now more formally linked to EuroHPC: ETP4HPC [42][43] and BDVA [44] are the private members of the JU.

The Big Data Value Association (BDVA) is an industry-driven international not–for-profit organisation with 200 members all over Europe composed of large, small, and medium-sized industries as well as research and user organizations. BDVA aims to develop the Innovation Ecosystem that will enable the data and Artificial Intelligence (AI) driven digital transformation in Europe delivering maximum economic and societal benefit, achieving and sustaining Europe's leadership on Big Data Value creation and AI.

ETP4HPC is an industry-led think tank composed of European HPC technology stakeholders: technology suppliers, vendors, research centers, independent software vendors (ISVs) and end users, with almost 100 members at the beginning of 2019.

Both associations were private partners of contractual Public Private Partnerships (cPPP) under Horizon 2020, resp. on HPC [45] and Big Data [46]. The two associations have developed their own Strategic Research and Innovation Agendas for several years, with already strong cooperation and joint efforts on technology and usage visions – where HPC and compute power meets Big Data, and, increasingly now, AI and IoT. These SRIAs served as recommendations and input for H2020 HPC and/or Big data calls. In particular ETP4HPC SRIA served as a technical reference for the FETHPC - HPC technology - calls, under the umbrella of HPC cPPP, between 2014 and 2017.

ETP4HPC became the European Commission's partner in the contractual Public-Private Partnership (cPPP) for High-Performance Computing at the end of 2013. The objectives of this cPPP were:

- Develop the next generation of HPC technologies, applications and systems towards exascale;
- Achieve excellence in HPC applications delivery and use.

Centres of Excellence in computing applications (CoEs) joined the cPPP governance in 2015.

The HPC cPPP will be terminated upon mutual agreement of both parties (the EC and ETP4HPC). From 2019 onward, the Joint Undertaking EuroHPC will take over, in particular regarding HPC R&I funding and steering. The exchange with the EC and the private partners will then be covered by ETP4HPC's participation of the Research and Innovation Advisory Board (RIAG) of the Joint Undertaking, in which BDVA will also participate.

3.1.4 Coordination and Support Actions

Horizon 2020 HPC-related Coordination and Support Actions that were launched in the FETHPC programme and thus strongly related to HPC technologies, are now running in their second phase (EXDCI-2 continues EXDCI; Eurolab4HPC2 continues Eurolab4HPC [47][48][49]).

EXCDI-2 continues the coordination of the HPC ecosystem with important enhancements with respect to EXDCI, so as to better address the convergence of big data, cloud and HPC. EXDCI-2 strategic objectives are a) Development and advocacy of a competitive European HPC Exascale Strategy and b) Coordination of the stakeholder community for European HPC at the Exascale. EXDCI-2 mobilizes the European HPC stakeholders through the joint action of PRACE and ETP4HPC. Beside continued support of roadmap efforts (ETP4HPC SRIA, PRACE Scientific Case), the organization of the annual EuroHPC Summit Week and other international actions, EXDCI-2 also works to increase the impact of the H2020 HPC research projects, by identifying synergies and supporting market acceptance of the results. In particular, one task of EXDCI2 Work Package 2 has been performing an analysis of the results of the FETHPC projects which began in September 2015 and, for most of them, finished in 2018. Their contribution to the position of Europe in HPC technologies is being assessed and a gap-analysis will be made.

Eurolab4HPC aims to consolidate European Research Excellence in Exascale HPC Systems - building connected and sustainable leadership in high-performance computing systems, by bringing together the different and leading performance orientated communities in Europe, working across all layers of the system stack, fueling new industries in HPC.

These continued actions remain fully consistent and useful in the context of EuroHPC and of its R&I Pillar in particular.

A new action for strengthened coordination of Centres of Excellence has also started end of 2018: FocusCoE [50] is described in the next section.

3.2 Applications: Centres of Excellence

Below is a quick reminder of Centres of Excellence for Computing Applications (CoEs) which represent the spearhead of EU strategy towards harnessing exascale infrastructures and to support scientific and industrial competitiveness and tackle societal challenges via computational methods.

Since this report is focused on HPC technologies, and this chapter focuses on the European HPC landscape, it is worth reminding here the manifold relationships between applications and technologies:

- Applications rely on technologies to provide them with computing power;
- Applications themselves encompass or invoke a lot of "software technology" (programming tools, software engineering, middleware…).

Not surprisingly, different levels of application/technology interactions can be found in the CoE projects, via co-design, benchmarking, optimization and other kinds of activities.

In 2015-2016, nine Centres of Excellence (CoEs) for computing applications have been selected, following an H2020 e-Infrastructures call, meant to help strengthen Europe's existing leadership in

HPC applications and cover important computational science areas [51]. The CoEs and their topics were:

- EoCoE - Energy oriented Centre of Excellence for computer applications;
- BioExcel - Centre of Excellence for Biomolecular Research;
- NoMaD - The Novel Materials Discovery Laboratory;
- MaX - Materials design at the eXascale;
- ESiWACE - Excellence in SImulation of Weather and Climate in Europe;
- E-CAM - E-infrastructure for software, training and consultancy in simulation & modelling;
- POP - Performance Optimisation and Productivity;
- COEGSS - Centre of Excellence for Global Systems Science;
- CompBioMed - started in 2016 - Computational Biomedicine.

In 2018, a second generation of CoEs has been selected [52], when most of 2015 CoEs came to their end. This includes continuations of most of the previous ones and centres dealing with new topics (ChEESE and EXCELLERAT). The second round of CoEs and their topics are:

- EoCoE-II - Energy oriented Centre of Excellence for computer applications;
- BioExcel-2 - Centre of Excellence for Biomolecular Research;
- MaX - Materials design at the eXascale;
- ESiWACE2 - Excellence in SImulation of Weather and Climate in Europe;
- POP2 - Performance Optimisation and Productivity;
- CompBioMed2 - Computational Biomedicine;
- HiDALGO – HPC and Big Data Technologies for Global Systems;
- EXCELLERAT – The European Centre of Excellence for Engineering Applications;
- ChEESE - Centre of Excellence for Exascale in Solid Earth;

A coordination and support action, FocusCoE [50] will be supporting the EU HPC CoEs to more effectively fulfil their role within the ecosystem. The project will create an effective platform for the CoEs to coordinate strategic directions and collaboration (addressing possible fragmentation of activities across the CoEs and coordinating interactions with the overall HPC ecosystem) and will provide support services for the CoEs in relation to both industrial outreach and promotion of their services and competences by acting as a focal point for users to discover those services.

3.3 Technologies: FETHPC, EPI

3.3.1 FETHPC

The aim of the H2020-FETHPC-2014 call for proposals "Towards Exascale High Performance Computing" was to attract projects that could achieve world-class extreme scale computing capabilities in platforms, technologies and applications. In total, 21 projects were selected and began in 2015. These included 19 Research and Innovation Actions (RIA):

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

2. ANTAREX - AutoTuning and Adaptivity appRoach for Energy efficient eXascale HPC systems;
3. ComPat - Computing Patterns for High Performance Multiscale Computing;
4. ECOSCALE - Energy-efficient Heterogeneous COmputing at exaSCALE;
5. ESCAPE- Energy-efficient SCalable Algorithms for weather Prediction at Exascale;
6. ExaFLOW - Enabling Exascale Fluid Dynamics Simulations;
7. ExaHyPe - An Exascale Hyperbolic PDE Engine;
8. ExaNest - European Exascale System Interconnect and Storage;
9. ExaNode - European Exascale Processor Memory Node Design;
10. ExCAPE - Exascale Compound Activity Prediction Engine;
11. EXTRA - Exploiting eXascale Technology with Reconfigurable Architectures;
12. greenFLASH - Green Flash, energy efficient HPC for real-time science;
13. INTERTWInE - Programming Model INTERoperability ToWards Exascale;
14. MANGO - Exploring Manycore Architectures for Next-GeneratiOn HPC systems;
15. MontBlanc-3 - European scalable and power efficient HPC platform based on low-power embedded technology;
16. NextGenIO - Next Generation I/O for Exascale - Project home page;
17. NLAFET - Parallel Numerical Linear Algebra for Future Extreme-Scale Systems;
18. READEX - Runtime Exploitation of Application Dynamism for Energy-efficient eXascale;
19. SAGE - Percipient StorAGe for Exascale Data Centric Computing.

In 2016, two more larger RIA project were selected under FETHPC-01-2016 call "Co-design of HPC systems and applications:

20. DEEP-EST – DEEP Extreme Scale Technologies;
21. EuroEXA - Co-designed Innovation and System for Resilient Exascale Computing in Europe.

Under the FET Proactive programme again, 13 FETHPC-02-2017 "Transition to Exascale Computing" projects were selected and began in 2018. These included 11 Research and Innovation Actions:

22. ASPIDE - exAScale ProgramIng models for extreme Data processing;
23. EPEEC - European joint Effort toward a Highly Productive Programming Environment for Heterogeneous Exascale Computing;
24. EPiGRAM-HS - Exascale Programming Models for Heterogeneous Systems;
25. ESCAPE-2 - Energy-efficient SCalable Algorithms for weather and climate Prediction at Exascale;
26. EXA2PRO - Enhancing Programmability and boosting Performance Portability for Exascale;
27. ExaQUte - EXAscale Quantification of Uncertainties for Technology and Science Simulation;
28. MAESTRO - Middleware for memory and data-awareness in workflows;
29. RECIPE - REliable power and time-ConstraInts-aware Predictive management of heterogeneous Exascale systems;

30. Sage2 - Percipient Storage for Exascale Data Centric Computing2;
31. VECMA - Verified Exascale Computing for Multiscale Applications;
32. VESTEC - Visual Exploration and Sampling Toolkit for Extreme Computing.

Like reminded and detailed in [37], FETHPC projects actually cover a diversity of hardware and software developments, dealing with a mix a building block or subsystem-level approaches and of more global system-level and architecture co-design and prototyping (involving 'pilot' applications in this latter case, such as in the DEEP [53] and MontBlanc [54] series of projects, and the EuroExa [55] project).

EXDCI2 Work Package 2 has performed an analysis of the results of the FETHPC projects launched in September 2015. All (19) projects participated in this survey, and a list of foreground IP they generated has been established. In total 171 results have come out of this first wave of HPC technology research projects. Around two-thirds of the results are software elements. In this field, we find a lot of interesting results to improve the HPC software stack with more energy efficiency, optimized parallel execution or better IO (input/output). Some of the software results are related to applications with the development of new implementations or optimizations for GPU or FPGA. In the area of hardware, the results span from core design to interconnect, with some FPGA- or ARM-based processor board developments. The hardware-related projects have developed 8 demonstrators, some of them being open to the whole HPC community for testing. Besides these software and hardware results, the projects have been active in proposing new APIs, defining benchmark suites and developing training sessions[1].

3.3.2 EPI

The European Processor Initiative (EPI) started in December 2018 as an H2020 project [56]. The overall aim of EPI is to develop IP owned in Europe for low-power microprocessors for the global market. Even though the focus is on delivering chips for HPC and in particular for Exascale supercomputers, the automotive industry for edge-HPC and broader datacenter market will also be targeted. For the implementation of this vision, the 23 European partners of the EPI consortium have signed a Framework Partnership Agreement (FPA). This FPA is currently planned to cover two Specific Grant Agreements (SGAs) with a total budget of €120M. These two SGAs will span a total period of four years, with SGA1 running from December 2018 to November 2021 and SGA2 from November 2020 to April 2022 and will allow the consortium to develop the technologies and tape out two revisions of the EPI 1st generation processor. Under SGA2, the initial technologies will be optimised and performances increased and revision 2 of the EPI 1st generation processor will be taped out. This chip will benefit from the experience gathered during SGA1, both in the development and usage of the EPI technology. Following this, EPI expects to develop the 2nd generation of the processor and the related technologies in subsequent SGAs which will target Exascale level systems and enable the derivation of chips for large-volume markets.

The EPI project will continue under the umbrella of EuroHPC.

[1] A more complete analysis will be available in a forthcoming EXDCI-2 report

3.4 Infrastructures: PPI4HPC, HBP

3.4.1 PPI4HPC

With the project PPI4HPC - Public Procurement of Innovations for High-Performance Computing [57] - four leading supercomputing centers in Europe formed a group of buyers to procure innovative High-Performance Computing (HPC) solutions. The project, co-funded by the European Commission under the call EINFRA-21-2017, marked the first initiative of this kind in the field of HPC and can be seen as a testbed to evaluate the usability of Public Procurement of Innovative Solutions (PPI) as instrument for future bigger joint pan-European operations launched in the field of the EuroHPC initiative.

A PPI is an approach defined by the European Commission (EC) in which several public procurers procure jointly innovative solutions and as such act as early adopters of innovative products or services that are newly arriving on the market (not widely commercially available yet). This approach aims at speeding up the availability of solutions meeting public sector requirements for large scale deployments. The EC provides co-funding to public procurers for stimulating innovation [58].

The public partners, namely BSC, CEA-GENCI, CINECA, and Juelich, worked together on coordinated roadmaps and on a joint procurement for providing HPC resources optimised to the needs of European scientists and engineers. The purpose of this procurement is, for each public procurer, to buy an innovative high-performance supercomputer and/or an innovative high-performance storage system that will be integrated into their computing center.

The PPI4HPC project aims at fostering the HPC ecosystem in Europe: by providing more computing, storage and data management resources to scientists and engineers; by strengthening R&I on HPC architectures and technologies that will incorporate innovative solutions that require strong relationship and possibly, collaboration between the procurers and the suppliers; and by sharing common topics of innovation and design of solutions that match the need of scientists and engineers in Europe and the public procures want to jointly foster.

The procurement was officially launched by publishing the contract notice on May 12th, 2018 [59], and each public partner will conduct their own competitive dialogue with the qualified candidates.

3.4.2 HBP: FENIX and ICEI project

The Human Brain Project (HBP) is a European research initiative to advance neuroscience and medicine and to create brain-inspired information technology [60]. It is funded by the EU's H2020 research funding program and it is one of the first two Future and Emerging Technologies (FET) Flagship projects. To achieve a comprehensive understanding of the brain, interdisciplinary expertise joining neuroscience, computer science, informatics, physics, and mathematics is needed. In this respect, one of the HBP primary objectives has been confirmed by the building of a European infrastructure for brain science. The creation of such an infrastructure requires research on system software, middleware, interactive computational steering, and visualization. In order to provide the necessary computational resources, storage, and networking a federation of the infrastructure of its HPC centers has started within the HBP.

To respond to such needs, which are also emerging from other science and engineering communities, five European Tier-0 supercomputer centers, namely BSC, CEA, CINECA, CSCS, and Juelich, agreed to join efforts in Fenix [61] to harmonise their service offerings within a federated e-infrastructure. Initial funding comes through the ICEI project that is executed under the HBP Framework Partnership Agreement.

The distinguishing characteristic of this new e-infrastructure is that data repositories and scalable supercomputing systems will be collocated within the same data center and well-integrated. The HBP will be the initial prime user of this research infrastructure but access to other domain researchers is also provided in collaboration with PRACE (with a first pilot phase started in PRACE Call #18).

The objectives of the ICEI project can be summarized as follows, encompassing significant R&D in software and hardware:

- Perform a coordinated procurement of equipment and related maintenance services, licences for software components, and R&D services for realizing elements of Fenix e-infrastructure;
- Design a generic e-infrastructure for the HBP, driven by its scientific use-cases and usable by other scientific communities;
- Build the e-infrastructure having the following key characteristics: Interactive Computing Services; Elastic access to scalable compute resources; and Federated data infrastructure;
- Establish a suitable e-infrastructure governance;
- Develop a resource allocation mechanism to provide resources to HBP users and European researchers at large; and
- Assist in the expansion of the e-infrastructure to other communities that provide additional resources.

4 Core Technologies and Components

4.1 Processors

4.1.1 x86_64 processors

4.1.1.1 Intel x86_64 processors

Due to the delays related to the 10nm manufacturing process, Intel's release calendar is not as predictable as it was in the past. While the company has announced a 3 step "Process-Architecture-Optimization" program in 2016, there have been releases in the non-server processor lines that extend into 2nd (e.g. Coffee Lake in Desktop line) and even 3rd (e.g. Whiskey Lake in Mobile line) optimization steps. In the Coffee Lake line, the only Xeon branded processors were under the "Xeon E" name, targeting the workstation market with a maximum of 6 cores. The successor, Cannon Lake, representing the Process step of the P-A-O cycle, with a die shrink to 10nm, has so far produced a single processor, i3-8121U, for the mobile market. Intel announced in July 2018 that volume production of this line will continue until late Q2 2019. The Cannon Lake series is still expected to be succeeded by Ice Lake (Architecture step) and Tiger Lake (Optimization step).

On the server market, the only significant release in the past year was Xeon Gold 6138p, a chip based on Skylake series with a built-in Intel Altera FPGA [62]. In July 2018, Intel officially announced the discontinuation of the Xeon Phi line [63], even if a Xeon Phi based system is still ranked 6th in the Top500 list [1]. Intel will, therefore, focus in the near future on its Xeon line for the HPC market, with the launch of Xeon AP/SP[2] lines to cover different market requirements. Some of the technologies employed in the Xeon Phi line, such as HBM on-chip, are expected to be transferred onto these lines. Further hardware innovations for the HPC market may also come from the DOE-funded joint project between Intel-Cray and Argonne, called Aurora21, targeting exascale performance in 2021 [64].

4.1.1.1.1 Intel Cascade Lake Scalable Processor

In the server-oriented Xeon line, the next processor release will be under the name Cascade Lake, and will come as a drop-in upgrade still based on the Purley platform. This represents the optimization step of the P-A-O fabrication execution, an enhancement of the 14nm process (denoted as 14nm++ manufacture). Similar performance compared to the previous Skylake core has to be expected. In the Cascade Lake SP line, the density will be comparable to the previous processors, with core count up to 28 cores per chip available.

In terms of memory support, these processors will be compatible with persistent memory chips, i.e. DDR-T/Optane DIMMs based on 3D XPoint technology (Apache Pass) [65], with support of up to 3 TB per socket. A 128 Gb/s memory bandwidth is provided through 6 memory channels. The cache hierarchy includes 1.375 MB per core 11-way set associative L3 cache (shared by all cores, non-inclusive), 1 MB per core 16-way set associative L2 cache (inclusive) and 32 KB per core 8-way set associative L1D cache (competitively shared by threads/cores) and 32 KB per core

[2] Advanced Performance and Scalable processor, respectively.

8-way associative L1I cache (competitively shared by the threads/cores) non inclusive. Finally, the chip is equipped with 48 PCI gen 3 lanes.

Furthermore, this will be the first series of Intel processors featuring hardware mitigations for CVE-2017-5715 (Spectre, Variant 2), CVE-2017-5754 (Meltdown, Variant 3) and CVE-2018-3620/CVE-2018-3646 (L1 Terminal Fault) vulnerabilities. Extensions to the instruction set will be limited to AVX-512 Vector Neural Network Instructions (AVX512_VNNI), which are claimed to provide a 3x and 2x performance improvement for Int-8 and Int-16 arithmetic precision respectively [66] for AI inference workload.

4.1.1.1.2 Intel Cascade Lake Advanced Performance

Besides the SP line, the Cascade lake series of processors will be including new processors named Cascade Lake Advanced Performance (Cascade Lake AP), announced by Intel at the end of 2018. It is expected to hit the market in mid-2019. The chip will feature a so-called Multi-Chip-Package design, where multiple dyes are glued together via a high-speed interconnect [67].

This chip is expected to be denser than the Cascade Lake SP, with a core count of up to 48 cores per socket and still based on a 14nm process. The development of these processors targets larger memory bandwidth workloads, featuring up to 12 DDR4 memory channels.

This CPU release should be the last stopgap processor before Intel 10nm architecture.

4.1.1.1.3 Intel Future (Cooper Lake and Ice Lake)

At the time of writing, Intel is planning a further successor to the Cascade Lake in the Xeon line still based on the 14nm fabrication process with the name Cooper Lake [68]. Information on the advancements that will be featured in this architecture are limited, but support for BFLOAT16 for improving AI application performance is expected.

After that, Ice Lake SP is expected to hit the market in 2020 [69] and will be the first Intel processor with a 10 nm architecture (Process of the P-A-O cycle). Few hardware details are available to the public so far: it is expected to have a higher core count than Cascade Lake SP, it will have AVX512 instructions and a core frequency comparable with previous processors of the Xeon line. The architecture will provide support for PCIe4.

4.1.1.2 AMD x86_64 processors

AMD targets the server market with the EPYC line of processors, codenamed Naples. This CPU is based on the Zen architecture with a 14 nm manufacture process. This is AMD's first step of a 3-year plan for the datacenter market, that will see the release of successor "Rome" based on a 7nm process and finally "Milan" (7nm+) in 2020 [70].

4.1.1.2.1 AMD EPYC (based on Zen)

AMD EPYC processors line (7 series) was launched in 2017 and it is based on a quad-die design. The smallest building block is a group of four ZEN cores, called CCX (CPU complex). Each core has its own 2 MB of L3 cache and shares the remaining 6 MB with slightly more latency. In the

EPYC chip, two CCXs are assembled together with a custom fabric[3] and constitute a Zeppelin dye. Finally, four Zeppelin dyes are tied together in an MCM model (Multi-Chip-Module). This design implies a latency penalty accessing data in L3 cache from a core of a different CCX [71]. The top processor of EPYC 7 line sees has a core count of 32 cores with a base frequency set to 2.2 GHz that increases to 2.7 when all cores are active. The chip is also equipped with 8 memory channels DDR4, for a theoretical bandwidth of 159 GB/s and a maximum of 2 TB of memory.

In 2019 the EPYC line of processors will be enriched with new higher frequency processors targeting latency workloads (e.g. EPYC 7371). In this particular configuration, the core count will drop to 16 cores, with a range of frequency operating from 3.1 GHz to 3.6 GHz and up to 3.8 GHz with only one core active.

It is also to be noted that AMD processors are not affected by the Meltdown vulnerability.

4.1.1.2.2 AMD 2nd Gen Ryzen and Threadripper (based on Zen+)

For the mobile and desktop market, AMD introduced the Zen+ architecture in April 2018, under the name of 2nd generation Ryzen and 2nd generation Threadripper. Produced using GLobal Foundries' 12nm manufacturing process, these processors provide a maximum of 32 cores in a single chip, as in Zen microarchitecture, but enable higher clock speeds and lower power consumption. Compared to the older 14nm fabrication process, up to 15% higher density or 10% higher performance was expected. The Ryzen series features a 10% frequency increase at the same power budget. However, as no die shrink is involved, not all of the opportunities for optimization were taken, in similarity to the Tick step of Intel's former Tick-Tock strategy. This is also reflected on the software stack, as no new compiler flags were added respect to the 1st Zen generation architecture. The top of the line, the 2990WX model, features 32 cores, operating at 3.0 GHz, with a boost frequency of 4.2 GHz.

Key changes compared to the preceding Zen architecture, apart from the new manufacturing process and clock frequency boost, are an improvement in the cache prefetch, lower L2 latency (12 cycles down from 17 cycles) and higher DDR4 transfer rate (2933 MT/s up from 2666 MT/s). The cache hierarchy consists of 8 MB per 4 cores 16-way set associative L3 cache (shared by all cores, non-inclusive), 512 KB per core 8-way set associative L2 cache (inclusive) and 32KB per core 8-way set associative L1D cache / 64KB per core 8-way associative L1I caches.

4.1.1.2.3 AMD EPYC - Rome (based on Zen 2)

The successor of the current EPYC processor - codenamed Rome - is expected to hit the market in mid 2019. It will be featured in one of the largest supercomputers in Europe, hosted at HLRS and provided by HPE [72] and capable of a theoretical peak of 24 PFlop/s. Rome, based on the Zen 2 microarchitecture, is expected to be the first server processor with a 7nm[4] process. This is the result of a long-term strategy that AMD started in 2015 to tackle the HPC market with the Zen microarchitecture. The upcoming Rome processor is claimed by AMD to provide 4x performance compared to the predecessor EPYC Naples, by doubling the core count - from 32 to 64 cores - and

[3] AMD infinity fabric.
[4] It should be noted here that AMD relies on TSMC 7nm shrink that should compare to Intel's 10nm shrink.

by doubling the vector floating point width, from 128 to 256 bits. However, the number of lanes connecting to the memory is limited to 8, compared to 12 lanes in Intel's offerings.

Rome is claimed to be immune to the Spectre vulnerability due to new on-chip mitigations and, as the predecessor, to be immune to the Meltdown vulnerability.

4.1.1.2.4 AMD Future (based on Zen 3, 4, 5)

At AMD's Tech Day in February 2017, a future iteration of the Zen architecture, Zen 3 was announced. Later that year, Zen 3 was confirmed to be based on the 7nm+ manufacturing process. A further key improvement will be that this architecture is expected to support up to 4-way SMP, instead of 2-way SMP of current AMD processors.

In late 2018, during different events, Zen 4[5] and Zen 5[6] iterations were also mentioned by AMD, confirming the interest of the company in the datacenter market.

4.1.2 ARM processors

In recent years an increasing number of server-class, 64-bit processors based on ARM technology have become available. These have reached a performance level that is on par with other widely used processors for supercomputers. The successful listing of the Astra system, a supercomputer with 125,328 ARM cores at Sandia National Laboratories (US) on position #204 of the November 2018 Top500 list supports this observation.

In the following, we discuss some selected processors, which may be considered most promising for HPC from today's perspective[7]. All have in common that they support the Armv8 ISA and support out-of-order scheduling of instructions. Furthermore, all are based on custom micro-architectures, i.e. not on micro-architectures provided by ARM. For a comparison of performance figures, see Table 7.

The ThunderX2 processor [75] was originally developed by Broadcom. Meanwhile, this processor is a Marvell product, which became generally available early 2018. This is the first ARM-based processor, which has been used for larger HPC systems, including Astra (US) and Isambard (UK).

Early 2019, Huawei officially released its most recent ARM server processor, which is now branded as Kunpeng 920 [76][77]. It features, compared to ThunderX2 and A64FX, a significantly larger number of cores.

The A64FX processor [78] has been designed by Fujitsu in collaboration with RIKEN for the Post-K supercomputer, which is planned to become operational in 2021. This processor is outstanding in several aspects. It is the first processor announced to support the SVE ISA with vectors of length 512 bits. Furthermore, to achieve a more balanced ratio between throughput of floating-point operations and memory bandwidth the high-bandwidth memory technology HBM is used (no support for DDR memory)

[5] AMD states this architecture to be in design completion phase.
[6] Sources seems to suggest a possible 5nm process for this architecture.
[7] It should, however, be noted that more server-class processors are being developed. Examples are the Xiaomi architecture, which is possibly the basis for one of the Chinese exascale prototypes [73], or the Ampere processor [74].

	Marvell ThunderX2	Huawei Kunpeng 920	Fujitsu A64FX
Number of cores	≤32	64	48+4
Clock frequency [GHz]	≤2.5	2.6	(not published)
Nominal throughput of double-precision floating-point operations [GFlop/s]	≤640	1331	>2700
Memory technology	DDR4, 8 channels	DDR4, 8 channels	HBM2
Nominal memory bandwidth [GByte/s]	170	204	1024
L1 data cache [kiByte/core]	32	64	64
L2 cache [kiByte/core]	256	512	8192/13
L3 cache [MB/core]	1	1	--
Process	14nm	7nm	7nm

Table 7: Comparison of different server-class ARM-based processors

With server-class ARM-based processors, which are suitable for HPC, becoming available, an increasing number of system integrators have started or announced ARM-based HPC systems:

- Atos-Bull has developed BullSequana X1310 blades with ThunderX2 processors that can be integrated into their BullSequana HPC rack system. The company announced a first larger installation comprising 276 nodes to be installed at CEA.
- Cray integrated the ThunderX2 processor in its XC50 solution and installed the Isambard system in Wales (UK), which claims to be the largest ARM-based supercomputer in Europe.
- Fujitsu is working on the Post-K supercomputer, which is planned to become operational in 2021 and is based on Fujitsu's A64FX processor.
- HPE developed a version of its Apollo 70 HPC platform with ThunderX2 processors. Based on this system the Astra supercomputer was deployed at Sandia Lab (US).

Along with the increasing number of ARM-based hardware solutions for HPC, the software ecosystem for ARM continues to develop and improve. Important software components include:

- Enterprise Linux distributions for ARM: Both Red Hat and SUSE provide full support for ARM-based versions of their enterprise Linux distributions RHEL and SLES, respectively.
- Parallel file systems: In November 2018, DDN announced to provide professional support for ARM-based clients for Lustre [79]. Also, BeeGFS continues to be available for ARM with support coming through ThinkParQ.
- Compilers: Support of compilers for ARM-based processors continuously improves. The main options are the freely available GNU and LLVM compiler suites. Furthermore, several commercial products are available, namely the ARM Compiler for HPC, Cray's CCE compiler and Fujitsu's compiler for the Post-K system.

An extensive collection of software components for ARM is readily available as part of the OpenHPC collection of packets for RHEL/CentOS and SLES[8].

[8]http://openhpc.community/

4.1.3 POWER processors

The POWER9 processor, which was already announced in 2016, became generally available in 2017. This processor is used for the supercomputers Summit and Sierra, which are listed on position #1 and #2 of the Top500 list as of November 2018.

There are two variants of POWER9 cores:

- POWER9 SMT4 cores: 4-way multithreaded core
- POWER9 SMT8 cores: 8-way multithreaded core, which is essentially a combination of 2 SMT4 cores

The SMT4 and SMT8 cores are organised in 4 and 8 slices, respectively, which are fed with instructions from a single scheduler. The instructions can be scheduled out-of-order. Each slice can perform a multiply-add in each clock cycle. The SMT4 cores are organised in pairs, which share an L2 cache as well as an L3 cache slice.

The POWER9 processors are available with up to 12 SMT8 or 24 SMT4 cores and available in the following variants:

- Scale-out variant: suitable for realising systems with a larger number of nodes.
- Scale-up variant: optimised for small NUMA systems.

	IBM POWER9 Monza Module
Number of cores	22 SMT4 cores
Clock frequency [GHz]	2.78 or 3.07 (turbo)
Nominal throughput of double-precision floating-point operations [GFlop/s]	≤540
Memory technology	DDR4-2666, 8 channels
Nominal memory bandwidth [GByte/s]	170
L1 data cache [kiByte/core]	32
L2 cache [kiByte/core]	512/2
L3 cache [MB/core]	10/2
Process	14nm

Table 8: IBM POWER9 processor as available for the 8335-GTW model of the IBM AC922 server as used for Summit [80][81].

A major difference between the scale-up and scale-out variants of the POWER9 processor is how memory is attached. The scale-up version of the processor attaches the memory through buffer chips similar as to preceding generations of the POWER processor. The scale-out version of the processor provides 8 DDR4 channels for direct attachment of memory DIMMs.

An outstanding feature of the POWER9 processor is the performance of its I/O interfaces. The scale-out variant of the processor comprises the following I/O interfaces [80]:

- PCIe GEN4: 48 lanes grouped in 3 sets of 16 lanes. The aggregate nominal bandwidth is 94.5 GByte/s per direction.
- 25G Link interface: 48 lanes grouped in 6 bricks with 8 lanes each. The aggregate nominal bandwidth is 147.7 GByte/s per direction.

The 25G Link interface can be used to attach NVIDIA V100 GPUs using the NVLink protocol. Up to 4 bricks can also be configured as OpenCAPI (Open Coherent Accelerator Processor Interface) interfaces. OpenCAPI is a new bus standard managed by the OpenCAPI Consortium. OpenCAPI allows for a cache-coherent attachment of devices like FPGAs as well as network, storage or other devices.

4.2 Highly parallel components/compute engines

4.2.1 GPUs

The following chapters describe the latest products of the two market leaders regarding GPUs for HPC computing.

4.2.1.1 AMD

AMD announced in June 2018 with the Radeon Instinct™ MI60 and MI50 the first 7nm Datacenter GPUs targeted to high performance computing, deep learning, rendering, and cloud computing with the following key features [82][83]:

- The AMD Radeon Instinct™ MI60 offering is 7.4 FP64 TFlop/s for HPC, 14.7 FP32 TFlop/s and 29.5 FP16 TFlop/s for deep learning training and 59 INT8 TOp/s for deep learning inferencing, while the MI50 has about 10 percent less performance (up to 6.7 TFlop/s FP64 peak performance).
- The MI60 offers 32 GB HBM2 ECC memory, while the MI50 comes with 16 GB. In both cases, AMD is claiming a record 1 TB/sec of memory bandwidth.
- Two Infinity Fabric™ Links per GPU deliver up to 200 GB/s of peer-to-peer bandwidth and the MI60 and MI50 are also the first GPUs that will be able to communicate with their CPU host over a PCIe 4.0 link, providing twice the bandwidth of PCIe 3.0.

4.2.1.2 NVIDIA

NVIDIA announced in September 2018 the Tesla T4 GPU, which incorporates the capabilities of the Turing™ architecture for higher deep learning inference via the hardware's enhanced INT8 and INT4 capabilities [84][85][86].

The Tesla T4 GPU has the following key features:

1. The T4 comes with 2,560 CUDA cores, along with 320 Turing Tensor Cores and a peak single precision floating point (FP32) performance of 8.1 TFlop/s, while its mixed precision (FP16/FP32) performance is 65 TFlop/s.
2. The T4 can deliver 130 Top/s for INT8 and 260 Top/s for INT4 and NVIDIA is claiming the T4 can process inference queries 20 to 40 times faster than a CPU, specifically, an Intel Xeon Gold 6140 processor.
3. The T4 PCIe card is equipped with 16 GB of GDDR6 memory, yielding over 320 GB/sec of bandwidth. Note that GDDR6 is something of a compromise here, inasmuch as the more performant HBM2 stacked memory that comes standard on the Tesla GPUs for training would have added additional cost. Unfortunately, GDDR6 memory tends to draw more

power than HBM2 on a capacity basis. That said, NVIDIA has managed to get the entire T4 package into 75 watts, which makes it suitable for the kinds of scale-out servers that this product is geared for.

4.2.1.3 Comparison - NVIDIA vs AMD

Performance tests have shown, that NVIDIA's new “Turing” T4 accelerator can beat the Radeon Instinct MI60 by about 19 percent in terms of images per second processed, because it has those Tensor Cores [87].

Until HPC-style codes start using Tensor Cores instead of floating-point units, these AMD GPUs are absolutely competitive with the Tesla GPUs for such work for his raw floating-point performance at single precision or double precision. But clearly, AMD needs to have a matrix multiply and accumulate unit to be competitive with NVIDIA on machine learning training.

On the software front, NVIDIA is clearly ahead of AMD in terms of availability and maturity of the software stack for HPC. The NVIDIA CUDA programming environment has now been available to end users for more than 10 years. In this period, multiple high performant libraries have been developed and now are at disposal to users that want to adapt their applications for GPU. However, given the interest of AMD in future exascale projects [88] and the relevant role that GPU devices may play in those projects, we can expect rapid progress of AMD in producing high quality software ecosystem for HPC to stay competitive in the field.

4.2.2 Others

Information on other similar technologies, such as FPGAs, the NEC Vector Engine and TPUs, can be found in the previous version of the Deliverable, D5.1 [5].

4.3 Memory technologies (volatile non-volatile)

4.3.1 Volatile memory technologies

The most used volatile memory is currently DDR4 which tops with DIMM size of 128GB running at 2933MT/s. According to Micron [88], one of the leading memory producers DDR4 will evolve up to 256GB DIMMs at 3200 MT/s. This is estimated in Q1/Q2 2020. In the meantime, the standardization committee JEDEC works on the finishing of the next generation memory specification called DDR5. While the specification is not yet finalized, Micron already plans to produce in 2020 DIMM starting at 64GB capacity running at 4800MT/s. Later increase of capacity and speed up to 6400MT/s is expected by Micron. Another leading memory producer, SK Hynix, announced [90] that a first DDR5 DIMM with 16GB capacity running at 5200MHz was already produced to demonstrate that the technology process is ready for mass production in 2020. This sample runs at 1.1V which means a 9% decrease in operating voltage compared to DDR4.

Another class of volatile memory technology important to HPC is High Bandwidth Memory, today usually implemented as stacked memory, allowing parallel access to multiple (up to 8) slices. The current generation (HBM2) is used especially on GPU accelerators like NVIDIA Volta and AMD Vega. It supports speeds up to 2GT/s and typically 4 slices (each 4GB) with 16GB capacity

providing a bandwidth of 256GB/s per slice. The standard update in 2018 by JEDEC [91] allows theoretically to go up to 12 stacks with each stack having 24GB and 2.4GT/s speed, providing bandwidth up to 307GB/s per slice. Just to put his into perspective with the mentioned DDR5 from SK Hynix, that provides 41.6GB/s. There are roadmaps for HBM3 which expect to increase density and capacity of slices, overall the number of slices and increase in bandwidth up to 512GB/s or more in 2020. But there are some issues with aspects like interposer space, thermals and others which are pushing the industry towards another type of high bandwidth memory – GDDR.

The GDDR is another type of volatile memory which is more similar to DDR rather than HBM but is mostly used in the same environments requiring higher bandwidth, e.g. graphics and HPC applications. Compared to DDR it operates at a slightly higher voltage (1.35V) but provides up to 16GT/s. Current products from SK Hynix used in the new RT line of NVIDIA GPUs run only at 14GT/s but with enough modules, the RT 2080TI card provides 616GB/s bandwidth for 11GB of GDDR6 RAM. According to Micron [92], the next generation of Achronix's FPGA will feature up to 8 GDDR6 DIMMs with up to 512GB/s bandwidth. This will offer a viable alternative (especially from the cost perspective) to HBM memories for HPC applications.

4.3.2 Non-Volatile memory technologies

The only generally available non-volatile memory technologies at the end of 2018 are the ones based on Intel and Micron’s 3D XPoint technology. According to Micron [93], the currently available options are DIMMs with 8GB@2133MT/s or 16GB@2666MT/s compatible with DDR4. The 32GB@2933MT/s should follow-up shortly in 2019. At SC18, Micron setup a live demo using the NVDIMM modules to show that in a virtualized environment they can achieve 4734MB/s bandwidth with 1183409 IOP/s while keeping only 1.465ms in latency to access the NVDIMMs. Micron stated that they can get 2x better performance on bare metal.

Intel announced their line of products, but still much information including latency, bandwidth and pricing is missing. This line of products is following the JEDEC NVDIMM-F specification, which means that the non-volatile part has to be used with standard volatile DIMMs. Overall this technology will definitively have an impact on the enterprise segment, optimal HPC applications (like check-point/restart) are yet to be tested and identified as promising or not.

4.4 Interconnect

4.4.1 Omni-Path, InfiniBand, BXI

InfiniBand technology is used in the top three fastest systems, including the new Linpack leader Summit at Oak Ridge National Laboratory and nearly 60 percent of the HPC category. Out of the total 500 grouping, Mellanox technology connects 216 systems with InfiniBand [94][95][96][97]. As of today, the 200Gb/s per port and 90ns port-to-port latency Mellanox HDR InfiniBand is the fastest interconnect fully available. This is combined with the availability of so-called Y cables (or splitter cables) which allow increasing the radix up to 80 ports per 1U if the bandwidth required for the endpoint is only 100Gb/s. Using Mellanox ConnectX-6 based HCAs, a single host can support up to two 200Gb/s ports. This needs full 16 lanes of PCIe gen 4, which is currently available only on the IBM POWER9 architecture. The x86 based systems having only PCIe gen 3 need to

utilize a Socket Direct Adapter [98], which is another 16 lanes PCIe adapter linked with the HCA using SAS cable and so provides the extra needed bandwidth for 200Gb/s connectivity to the network. In 2 socket systems, the same adapter can also be used to directly connect both CPUs to the network. This bypasses the inter CPU routing over QPI/UPI (Intel), Infinity Fabric (AMD) or CCPI2 (ARM).

Intel's Omni-Path Architecture (OPA) is the interconnect on 39 machines out of the top 500. Currently, it supports only 100Gb/s single port adapters with switches providing 48 ports per 1U. The next generation should support 200Gb/s per port requiring 16 lanes of PCIe gen 4 and switches with 64 ports in 1U. The expected availability is in late 2019, although many shifts in time provide this technology [99].

Both technologies are the prime one used today in HPC with a good support in the next generations of system solutions, like the new Cray Shasta (IB, OPA) [100][101], ATOS/Bull Sequana XH2000 (IB) [102][103] and HPE SGI 8600 next gen (IB, OPA) [104]. While for Cray the first choice of interconnect for HPC will be the new Slingshot, Atos will probably try to propagate more their BXI when it's suitable (e.g. from the price perspective), since it only provides 100Gb/s.

Apart from Infiniband and OPA, out of the top 500 there are 48 systems that make use of Cray (Aries/Gemini) technology and 26 systems that are using some other flavor of custom/proprietary interconnect (BlueGene, Fujitsu Tofu, Bull BXI, NUDT's TH Express-2, etc) [94][95][96][97].

4.4.2 Spectrum Ethernet, Slingshot

Ethernet seems now to be the leading networking technology with clear roadmap defining speeds up to 1.6Tb/s and offering 400Gb/s per single port [105]. This is caused mainly by the demand and supporting financial investments from the big cloud providers/hyperscalers (US and China) [106]. It starts to influence the capabilities of the InfiniBand through Mellanox and their VPI adapters, supporting both IB and Ethernet, where for more than a year the incomes coming from the Ethernet are playing a major part in the total volume [107].

The Mellanox Spectrum SN3510 Open Ethernet Switch using Spectrum™-2 ASIC supporting 50G PAM-4 signalling provides bidirectional switching capacity of up to 12.8Tb/s with a landmark 8.33Bpps packet processing rate and a port speeds up to 400Gb/s [108]. The six 400Gb/s can be either used to connect other switches or using split cables to connect to 2x200GB/s, 4x100Gb/s or 8x50Gb/s ports (like adapters in the servers). Currently, only Direct Attached Copper cables are available as split cables [109].

But not only Mellanox sees the potential of having an Ethernet-based HPC interconnect. Cray will use for his new Shasta platform the full-on Ethernet compatible Slingshot interconnect. Cray has designed an Ethernet protocol that has all of the hallmarks of a good HPC interconnect: smaller packets, smaller packet headers, reliable hardware delivery across the link, credit-based flow control that gives the characteristics of an HPC network. Slingshot has a crop of new features aimed at data-centric HPC and AI workloads. It starts with extremely high bandwidth: 12.8 Tb/s/dir per switch, from 64 200Gb/s ports [110][111]. Slingshot will fully support the DragonFly topology developed for Aries interconnect and will allow connecting directly to third-party Ethernet-based storage devices and to datacenter Ethernet networks, thus very much simplify the integration with the rest of the datacenter.

Slingshot, following Aries and Intel Omni Path will rely on libfabric: a software component which replaces the libibverbs which standardized RDMA for InfiniBand. The libfabric was done as a new approach to avoid reimplementation of IB functionalities and dependencies for Ethernet which doesn't provide them. It's also used by Cisco, a well-known Ethernet vendor, who's focusing more on the hyperscalers and cloud computing rather than on pure HPC. One of the reasons is the fact, that the current Ethernet is still lacking the low latency compared to InfiniBand. The latest Cisco Virtual Interface Card have on the link level 1570 ns compared to the 600ns on the latest ConnectX-6 VPI cards from Mellanox. On the port-to-port level of a single switch, Ethernet provides 450ns (ARISTA 7260CX3-64 switch with 100Gb/s ports) where the Mellanox InfiniBand provides 90ns (QM8700 switch with 200Gb/s ports). ARISTA also offers the 7060PX4-32 switch with 32 ports at 400Gb/s with the port-to-port latency of 700ns [112][113]. It's worth to mention, that in 1U the 7060PX4-32 switch can connect up to 128 100Gb/s devices (using split cables), which tops even the Mellanox with 80 devices per 1U.

4.4.3 Gen-Z, EXTOLL, Dolphin

The Extoll interconnect started with a switch-less design where the host cards build a 3D Torus network. The new addition to the Extoll family represents a switch called Fabri3 (pronounced Fabri Cube) [114]. This is internally a 3D mesh with 2*4*8 dimensions where the added value as Network Attached Memory or FPGA acceleration can be used allowing for high-speed processing of data in the switch. It also centralizes nicely the cabling topology. The host bandwidth to/from the cards can be up to 8.2Tb/s where internal switching capacity tops at 14.4Tb/s.

The Gen-Z interconnect tries to address both intra-node connections as well as inter node ones. The protocol is defined for some time, but now a set of connectors, cables and media converters from different vendors from the consortium implementing the protocol in real products starts to show. A live demo was presented at SC18 showing the connection of DDR memory, as well as NVMe over the Gen-Z protocol featured hardware where an additional latency of only 200ns was measured compared to the native connection of DDR memory to CPU. It's expected to go up to 400-500ns for real life big complex systems [115]. The intra-node connections to FPGA accelerators are now implemented on top of the PCIe on the x86 systems but a native Gen-Z implementation is expected later probably with the AMD Rome CPUs. For the inter-node connections, QSFP is currently used with expectations to reuse the technology coming from the 400 Gigabit Ethernet.

The Dolphin interconnect uses PCIe lanes and an external switch to create an "intelligent PCIe network" amongst up to 64 nodes [116]. Using copper cables with up to 16 PCIe lanes it can reach 9m from host to switch. Using fiber cables up to 100m connections can be used. The network can offer PCIe multicast capabilities for features like reflective memory [117] as well as IP-over-PCIe for more general HPC and AI workloads.

5 Data Storage and Data Management – Technologies and Components

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

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

The data infrastructure importance is growing up with the increased data capacity. It becomes more and more important in both HPC and cloud environments. It is obvious that hardware has to be supported by software solutions. We can observe user requirements, which are changing the cutting edge of a common e-Infrastructure, including parallel and distributed computing

5.1 Offline Storage (tapes)

5.1.1 IBM tape storage

The TS4500 tape library remains the IBM mainstay for large tape storage deployments. Currently, it offers the choice of LTO and 3592 technology in the same library, but the drive mounts and cartridge slots are not interchangeable and need to be decided on a frame by frame basis. In total, the TS4500 supports 18 frames.

Newer tapes drives have been introduced in more rapid succession in recent years, the 3592 series has received upgrades in both 2017 and 2018, breaking the previous three-year cadence. TS1150 (gen 5) was introduced in 2014 along with the JD media format (10 TB uncompressed) and received the TS155 (enhanced gen 5) upgrade in 2017. The TS1155 increased the amount of data that could be stored on JD media to 15 TB and also introduced ethernet connectivity as an option instead of Fibre Channel.

The TS1160 (gen 6) drive was introduced in Q4 2018 and introduces a new media format (JE) supporting 20 TB of uncompressed data per cartridge. Connectivity options now include 16 Gbit/s FC and 25 Gbit/s Ethernet.

5.1.2 Spectra Logic tape storage

Spectra Logic continues their focus on the high-performance tape storage market with proprietary features such as TAOS (see section 5.2.1 below) for LTO performance. Libraries produced by Spectra Logic uses tape drives from IBM, see section 5.1.1 for general information.

Support for TS1160 drives has been announced, bringing up the total capacity of the high-end TFinity library to almost an Exabyte of uncompressed data. Oracle T10k drives are also still supported in the TFinity library making it the only supplier to provide support for all current tape technologies in the same library.

On the T950 library LTO and TS11xx drives are supported, with TS1160 support coming in 2019. A lower cost version of the T950, named T950v, was introduced in 2018. One of the options to lower the cost is the possibility of using half-height LTO drives, which are cheaper but cannot use the full bandwidth of LTO-8.

5.1.3 Oracle StorageTek tape storage

With the end of development on the T10k media format, Oracle is using LTO drives for future capacity upgrades. The SL3000 library has been replaced by the SL4000 library in the product portfolio. One major change is the increased capacity compared to the SL3000 due configurations with more slots being possible, thus filling the previously rather large gap to the SL8500. Both libraries support LTO-8 and T10000D drives for storage density.

5.1.4 Quantum tape storage

Quantum is the only library vendor focusing entirely on LTO technology. For large scale tape deployments, the Scalar i6000 library is the only option that scales beyond one rack. Supported drive options are LTO-6, LTO-7 and LTO-8.

Another Quantum product is the StorNext Storage Manager which integrates support for data life cycle management using the i6000.

5.1.5 Performance optimizations for tapes

In recent years, the subject of how to efficiently utilize tape as a storage medium has seen renewed research interest. Industry trends favoured disk storage and tape became a mostly write-only medium used for disaster recovery and air-gapped copies. Historically, tape has also been used as on-line storage, especially early in the history of computing, and research on how to reduce latency of read operations was an active field. With recent slowdowns in the growth of hard drive sizes, combined with large amounts of data that needs to be stored for long times, using tape for colder but still on-line data has become more attractive.

Market segmentation has created a division between LTO based systems for price optimized deployments and the so-called “enterprise tape” technologies are used for performance optimized tape systems, but after Oracles decision to stop developing the T10k technology only IBM 3592 remains in the enterprise segment. The resulting lack of competition in the tape drive space has driven a renewed interest for research into how existing drives can be used more efficiently, and more specifically how to increase the performance of LTO based libraries.

Main differentiators between LTO and Enterprise tape have been capacity, streaming bandwidth and access latency. As seen in Table 9, the bandwidth is almost equal today, and the capacity gap has been shrinking with roadmaps converging in the future. Today latency for reading files is the main factor for choosing 3592 drives.

Tape technology	Introduced	Uncompressed capacity	Bandwidth
LTO-8	2017Q4	12 TB	360 MB/s
3592 Gen 6 (TS1160)	2018Q4	20 TB	400 MB/s

Table 9: LTO and Enterprise tape capacity and bandwidth

Data is stored in a linear serpentine pattern on current tape media, so when reading multiple non-contiguously placed files the optimal order is often not the order in which they were written. Doing a linear recall of files may result in long seek times where the tape is winding back and forth. The 3592 drives implement Recommended Access Order (RAO) whereby applications can query the tape drive for the order it should request files to be read. Searching for a file on tape uses a tape directory for high-speed winding to the start of the right section of tape, this has a higher resolution on 3592 than LTO (64 vs 2). Taken together this decreases the recall times for enterprise tape systems.

Spectra Logic has a proprietary implementation named Time-Based Access Order System (TAOS) [118] introduced for their high-end tape libraries (T950 and TFinity) in mid-2018. In effect this implements functionality similar to 3592 RAO on the library level for LTO-7+ tape drives by adding a “TAOS processor” to the drive sled. This extra processor appears as a separate LUN on the same SCSI ID and can be queried using the same SCSI commands as used for RAO. Application support is required to use this with a limited set of applications supporting it at the moment.

CERN has based their tape libraries on 3592 and T10k drive technologies, and after the demise of T10k, they have been investigating using LTO more efficiently. They have previously implemented RAO support [119] in their tape archiving software and presented [120] at HEPiX fall 2018 on tests on their own client-side implementation of something similar to RAO, but queries the drive for the tape head position during usage and builds its own map. This research is based on work done in the late 90s, so it is not a new problem, but one given less consideration when the trend was that tape usage declined. Plans are to implement this in the CERN software during 2019, in this case CERN controls the entire stack and does not have to wait for vendor support.

5.1.6 LTO media patent dispute

Fujifilm and Sony are the manufacturers of LTO cartridges today, and they have been embroiled in a dispute since late 2016 over patents related to LTO media. In March 2018 Fujifilm received a “Final Determination” from the US International Trade Commission barring Sony from importing LTO-8 media. This has led to a shortage of LTO-8 media worldwide, making it very hard to buy the highest capacity of LTO media.

LTO-8 drives are available, it is only the media availability that is affected by this dispute. It does make it hard to take advantage of the LTO-8 features, with the exception of using LTO-7 media in M8 mode.

5.2 Online storage (disk and flash)

This section focus on a limited number of suppliers and technologies that have evolved significantly compared to the previous deliverable.

5.2.1 DDN

DDN is the main provider of storage solutions for HPC: storage solutions from DataDirect Networks (DDN) are installed in over 2/3 of the fastest supercomputer environments.

DDN provides complete, end-to-end solutions across performance, archive and cloud for data-intensive, global organizations and Universities around the world.

The current products from DDN have new advancements in block and file storage, scale-out SSD and NVMe and solutions fully integrated and optimized for AI (artificial intelligence) and DL (deep learning) workloads [123][124][125].

These products include:

- DDN SFA18k: The SFA18K is a scalable hybrid storage solution that aims at accelerating processors, embeds file systems, and optimizes containers. It provides up to 3.2 million IOPs and 90GB/sec from a single 4U appliance with a mix of flash (NVMe and SAS) devices and spinning drives. It offers a capacity up to 13 PB in a single rack.

Figure 27: The 4U appliance of DDN SFA18k

- DDN A³I (Accelerated, Any-Scale AI) solutions with NVIDIA DGX-1 are engineered from the ground-up for AI-enabled data center to accelerate AI applications and streamline DL workflows, deliver faster performance, effortless scale, and simplified operations. The AI200/AI400 and AI7990 support a scale-out model with solutions starting at a few TBs yet scalable to 10s of PBs with up to 360TB of scale-out NVMe capacity per AI200/AI400 appliances, or 5.4PB of hybrid storage in the AI7990.
- DN ES7990 (ExaScale) and GS7990 (GridScale) are the integrated file system solutions for technical computing, Big Data and AI markets. The ES7990 and GS7990 provide sequential read/write performance with up 20/16 GB/s and up to 700,000 IOPs per appliance.
- DDN IME (Infinite Memory Engine) is an SSD and NVMe scale-out, flash-native data cache that alleviates the challenges and inefficiencies caused by I/O scale and bottlenecks. DDN offers a software-only solution for those environments with IT-based limitations on supported hardware and the integrated appliances IME140 (1U) and IME240 (2U) with up to 23 NVMe SSD drives, which provides with up to read/write performance with up 20/11 GB/s and more than 1M file IOPs.

- DDN DataFlow performs backup, migration, and archiving at-scale, from a single pane of glass. Designed for large volumes of data to ensure constant availability and security.
- DDN ES14KX® Lustre Appliance provides a performance with up to 50GB/s per Base Enclosure, a capacity with up to 17.5PB and scales to 100s of PB [126].

DDNs trends regarding 2019 are the following [127]:

- The emergence of large-scale AI and Machine Learning deployments.
- The advancement of granular data management capabilities for at-scale data systems and private clouds.
- The move toward cloud-like data management models for on-premise deployments with transparent mobility to the public cloud.
- Accelerated adoption of at-scale Flash deployments and the ascendency of NVMe, which will be the default media for tier-1 applications for the low latency, high IOPs and density, but NVMeOF will continue to lag as a networking standard as other more established RDMA networks like InfiniBand and RoCE continue to thrive and meet performance demands.

5.2.2 Seagate

Seagate, a subsidiary of Cray since 2017, as one of the leaders in data storage solutions, launched on June 21, 2018, a wide range of 14TB hard drives, which includes the Exos X14 drives for hyperscale data centers. These helium-based drives offer enhanced areal density to deliver higher storage capabilities in a compact 3.5-inch form factor and deliver 40% more petabytes per rack compared to Exos 10TB drives, while maintaining the same small footprint, a 10% reduction in weight versus air nearline drives, flexible formatting for wider integration options and support for a greater number of workloads [128][129][130].

In August 2018, Seagate showcased its latest enterprise solid-state drives, which include the 8TB Nytro XP7200 NVMe (10 GB per second), the 60TB SAS, and its 2TB Nytro XM1440 M.2 NVMe (high density M.2) [131].

For the affordability of flash, solid-state drives in general and the NVMe technology are becoming more widespread and part of a lot of storage solutions from vendors like DDN and Dell EMC [132].

5.2.3 HAMR and MAMR drives

Current hard-drives mainly use PMR (perpendicular magnetic recording) technology. The energy required to reverse the magnetization of a magnetic region is proportional to the size of the magnetic region and the magnetic coercivity of the material. The larger the magnetic region is and the higher the magnetic coercivity of the material, the more stable the medium is. Thus, there is a minimum size for a magnetic region at a given temperature and coercivity. If it is any smaller it is likely to be spontaneously de-magnetized by local thermal fluctuations. Perpendicular recording uses higher coercivity materials because the head's write field penetrates the medium more efficiently in the perpendicular geometry. The popular explanation for the advantage of perpendicular recording is that it achieves higher storage densities by aligning the poles of the magnetic elements, which represent bits, perpendicularly to the surface of the disk platter.

PMR drives have passed 1Tbit/in^2, with Toshiba's single platter MQ04 2.5-inch disk drive and its 1TB capacity as an example. At the current stage, PMR seems to have many issues, like the tendency of magnetic polarity and in hence bit-flipping between neighbour data areas, as they become more unstable with size reduction and bit-density.

To overcome these problems, two technologies were announced by the two biggest hard-drive manufacturers: HAMR (Heat-Assisted Magnetic Recording) technology by Seagate, and MAMR (Microwave-Assisted Magnetic Recording) by Western Digital.

Seagate uses a laser-assisted (with Seagate-developed plasmonic near-field transducer NFT) head to pre-heat a data region up to 400C for a very short time (~1 nanosecond) before writing the data on the drive surface, and effects on stable bit recording. With in-lab reliability tests a single HAMR head passed continuous 6000 hours, 3.2 PB reliable transfer test.

Seagate is also actually developing performance-optimised HAMR drives with MACH.2 multi-actuator technology – two read/write heads per platter – and capacity-optimised drives with shingled magnetic recording (SMR). Seagate claims that they can achieve up to 480MB/s sustained throughput on a single hard-drive equipped with MACH.2.

The First HAMR EXOS X14 14TB drive prototype was presented on CES2019 on January 2019 in Las Vegas [133] and it was announced that first HAMR EXOS drives will be shipped to customers in 2019 for testing. As for today's announcements Seagate produced about 40.000 HAMR 20+TB drives for clients testing purposes, built on the same assembly line as current products. They test-prove 2PB single HAMR head data transfer and achieved 2Pb/inch and 30% annual density growth in the past 9 years of technology development and announced ~2.5M MTBF in new drives.

Western Digital MAMR technology as manufacturer's claims, overcome HAMR issues by adding microwaves to the write head, using a spin-torque oscillator (STO) to generate magnetic domains. Electrons in a magnetised area have a spin state, tending to spin one way or another. By applying microwaves at the right frequency, a resonance effect can alter the spin state and make it easier for the write head's electrical field to alter the magnetic polarity of the domain. WD claims that they can reach a 4Tbit/in^2 areal density over time using MAMR technology, with a 15% compound annual growth rate (CAGR) in capacity (Table 10).

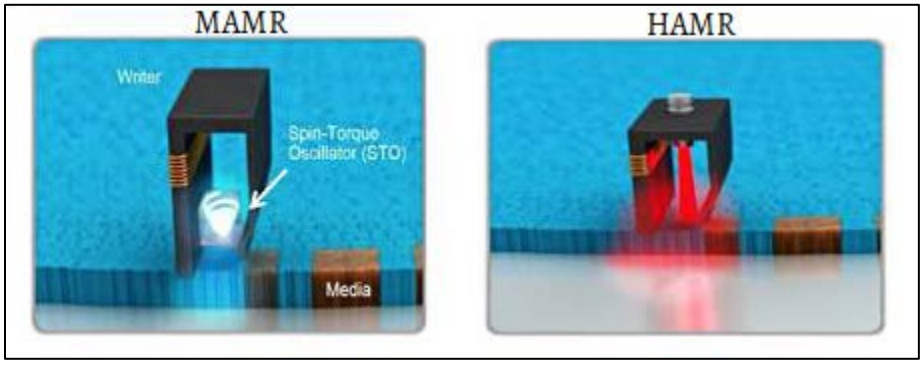

Figure 28: The presentation of MAMR and HAMR technologies

Description	MAMR	HAMR
Added cost	"Damascus" technology head	Laser, glass media, iron platinum coating
Complexity	Leverages current PMR technologies	New materials required
Reliability	High reliability	Test proven 3.2PB 6000 hours continuous single head read/write transfer
Manufacture-ability	MAMR product drive to be announced in 2019	40000 14TB+ EXOS drives produced, to be shipped to customers tests in 2019

Table 10: MAMR and HAMR comparison of technologies

HAMR/MAMR bits can be smaller than current PMR regions, resulting in a capacity increase. Seagate announced a roadmap of releasing 48TB by 2023 and doubling real density every 30 months, up to 100TB in HAMR drives 3.5-inch drives possible by 2025/2026 (Western Digital 100TB by 2030).

6 Overview of Vendor solutions/roadmaps

In this chapter, an overview of the solutions of the vendors in Top50 (see 2.1.4) and their corresponding roadmaps are presented, sorted in alphabetical order.

6.1 ATOS-BULL

ATOS technologies HPC branch has both BullSequana X supercomputers and X800/X550/X400 series cluster supercomputers. Atos/Bull currently has 22 supercomputers listed in the Top500 (Nov2018) [1].

6.1.1 BullSequana XH2000

The Atos HPC product line [102][151] got a new release on November 12, 2018 with the launch of the BullSequana XH2000 supercomputer. The BullSequana XH2000 is a hybrid supercomputer designed for running traditional high performance computing simulations with AI/deep learning workloads together on a single system. XH2000 is managed with SCS (Super Computer Suite) 5 which comes in two flavors; Single Island for up to 1,200 nodes and High End, which scales up to exascale configurations. Atos is also providing add-on software solutions for hybrid computing: Codex AI Suite, a toolbox for machine- and deep-learning frameworks and Extreme Factory hub for cloud integration to public and private cloud.

BullSequana XH2000 provides a system that supports the latest CPU and GPU processor and accelerator architectures, including Marvell ThunderX2 ARM, NVIDIA Volta V100, Intel Xeon processors and the latest AMD EPYC Rome. This is the first Atos supercomputer that supports AMD processors. The BullSequana XH2000 allows a choice of system interconnect technologies, including InfiniBand HDR and BXI (Bull eXascale Interconnect) and Fast Ethernet.

Atos currently lists four 1U blades in the XH2000 platform: the BullSequana X 1115, the GPU acceleration blade with a single node equipped with two Intel Xeon scalable processors and four NVIDIA V100 GPUs; the BullSequana X 1120, an Intel-only blade, three nodes, each of which is equipped with two Xeon scalable processors; the BullSequana X 1310, an ARM blade, three nodes, each equipped with two Marvell ThunderX2 CPUs; and the BullSequana X H2410, an AMD EPYC blade, three nodes, each equipped with two Zen 2 “Rome” processors. The AMD and ARM blades support up to two terabytes of memory.

The Open Sequana architecture makes BullSequana XH2000 compatible with future computing blade and interconnect technologies. It is exascale-ready, capable of processing 10^{18} operations per second by 2020. The BullSequana XH2000 is 100% water-cooled using Atos’ patented DLC (Direct Liquid Cooling) solution, which minimizes global energy consumption by using warm water up to 40°C inlet. Atos is claiming the XH2000 can provide a Power Usage Effectiveness (PUE) of “very close to 1,” which is the theoretical minimum. The BullSequana XH2000 is available from February 2019.

6.1.2 BullSequana X1000

BullSequana X1000 is the multi-petascale line for completely Direct Liquid Cooled (DLC) systems from Atos 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. A single compute cabinet holds up to 144 compute nodes and a hydraulic module to cool them. The switch cabinet contains level 1 DLC switches - BXI or InfiniBand EDR, level 2 Direct Liquid Cooled switches (BXI or EDR).

In the BullSequana 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.

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

6.1.3 BullSequana X400

BullSequana X400 series of rack-mounted servers designed for High-Performance Computing offer a balance between cost and efficiency. The BullSequana X440 E5 compute/service nodes integrate 4 2-way nodes within a rack-mounted 2U chassis. Equipped with new generation Intel Xeon Processors Scalable Family, these more cost-effective servers target usage as compute nodes within a cluster.

BullSequana X410 E5, dense GPU-accelerated compute node of 1U support up to four NVIDIA Tesla P100 GPUs, interconnected either via PCI-E direct connect or via NVIDIA NVLink. They also feature two Intel Xeon Scalable processors.

For service nodes in a cluster Bull offers both 1U and 2U servers. The BullSequana X430 E5 1U I/O node is a rack-mounted 1U mono socket server optimized to serve as an I/O node within a small to medium High-Performance Computing cluster. The BullSequana X430 E5 is a 2U rack-mounted 2-socket server, suitable as a service node, providing advanced connectivity features, extended storage options and redundancy features. X-server family also has X440 K5 model for Xeon Phi and X450 E5 graphics node for visualization needs.

6.1.4 BullSequana 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.

6.1.5 BullSequana 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.

6.2 Cray

Cray's Products have been divided between Computing, Storage and Analytics/AI. The computing product line has both XC series supercomputers, CS series cluster supercomputers and new Shasta supercomputers. The storage product line includes Cray ClusterStor storage solutions and Cray Datawarp I/O accelerators. On the analytics products line Cray has the Urika GX/XC analytics platform and software environments, the Cray graph engine and Urika-CS AI and analytics suite. Cray has 52 systems in the current Top500 (Nov. 2018) list.

6.2.1 Cray computing product line

6.2.1.1 Cray Shasta supercomputer

Cray released Shasta on Oct.30 2018 with a new high-speed interconnect called Slingshot. The Shasta system was first showcased at the SC18 conference in Dallas. Cray Shasta is an entirely new design, created to reach exascale performance for diverse workloads and processor architectures.

Cray Shasta can mix and match processor architectures (x86, ARM, GPUs) in the same system as well as system interconnects from Cray (Slingshot), Intel (Omni-Path) or Mellanox (InfiniBand). In Shasta, the system can have a heterogeneous mix from single to multi-socket processor nodes, one to 16 nodes per blade, GPUs, FPGAs and other forthcoming processor technologies, for example, AI specialized accelerators. Shasta can handle future's increasingly power-hungry processors with direct liquid cooling and supports W4-class warm water cooling. Cabinet cooling goes up to 250 kilowatts at first and increases up to 300 kilowatts per cabinet during the launch year, according to Cray. The Shasta architecture supports multiple cabinet types, a 19" air- or liquid-cooled, standard datacenter rack and a high-density, liquid-cooled rack designed to hold 64 compute blades with multiple processors per blade. Both options can scale to well over 100 cabinets.

The Cray-developed Slingshot interconnect will have up to 5× more bandwidth per node compared to existing XC-series and is designed for data-centric computing. Slingshot will feature Ethernet compatibility, advanced adaptive routing, congestion control and quality-of-service capabilities. Support for both IP-routed and remote memory operations will broaden the range of applications beyond traditional modelling and simulation. Quality-of-service and novel congestion management features shall limit the impact to critical workloads from system services, I/O traffic, and co-tenant workloads, to increase realized performance and limit performance variation. Reduction in the network diameter from five hops (in the current Cray XC generation) to three will reduce latency (300 nanoseconds per hop) and power while improving sustained bandwidth and reliability. Slingshot's 64-port switch with 200 Gbps ports (based on 50 Gbps signalling technology) provides 12.8 Tbps bandwidth per switch.

Shasta systems are expected to be commercially launched in late 2019. The first confirmed deal of Shasta has been announced with NERSC. The National Energy Research Scientific Computing

Center has chosen a Cray Shasta supercomputer for its NERSC-9 system, named "Perlmutter," in 2020. The system will feature AMD EPYC processors and NVIDIA GPUs offering a combined peak performance of ~100 PFlop/s. The program contract will feature a Shasta system with Cray ClusterStor storage.

6.2.1.2 Cray XC series supercomputer

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 TF per cabinet. Cray XC50 supercomputers can be configured up to hundreds of cabinets and upgraded to nearly 300 PF per system with CPU blades and over 500 PF 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 TF peak performance in a 24" cabinet with no requirement for liquid cooling or extra blower cabinets. Targeted for dedicated test, development, AI and analytics use cases, the air-cooled XC50 system has the same properties as the 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.

6.2.1.3 Cray CS series cluster supercomputers

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 is capable of scaling to over 27,000 compute nodes and 46 peak PFlop/s.

The CS400-LC system is a 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 TF of double-precision performance. For machine learning, where integer operations matter, a single CS-Storm 500GX server node can deliver up to 170 TOp/s (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 TF of double-precision performance. For machine learning, where integer operations matter, a single CS-Storm 500GX server node can deliver up to 170 TOp/s.

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.

6.2.2 Cray analytics and AI

The 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.

Cray also provides cluster systems in the CS product line. Cray CS500 system supports for 64-bit Intel Xeon Scalable processors and 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 peak PF.

6.3 Fujitsu

Fujitsu continues to be the main provider of high-end HPC solutions in Japan. Its K Computer, which was on position #1 of the November 2011 Top500 list and uses a custom SPARC processor, is today still on position #18. Fujitsu commercialised the K Computer technology in the PRIMEHPC FX10 and PRIMEHPC FX100 products. In the November 2018 edition of the Top500 list, 5 systems are based on PRIMEHPC FX100, which are all installed in Japan.

Meanwhile, Fujitsu deployed faster systems based on their commodity Primergy product line, which are based on x86 processors. With the Post-K system, Fujitsu is preparing for another generation of top-end supercomputers, which are also based on custom a processor and interconnect.

Primergy is Fujitsu's line of servers allowing for densely integrated x86 processors. For instance, the CX1640 M1 server is a half-width, 1U server for Intel KNL processors. With its support for liquid cooling, it allows integrating 8 such processors in 2U rack space by mounting 2 CX600 M1 chassis with 4 CX1640 M1 servers each from two sides of a rack. This commodity product has been used for the 25 PFlop/s system Oakforest-PACS, which is operated by the Joint Center for Advanced High-Performance Computing (JCAHPC) of the Universities Tokyo and Tsukuba and is listed at position #14 of the November 2018 Top500 list.

A similar modular server system was used for the ABCI system at Japan's National Institute of Advanced Industrial Science and Technology (AIST), which is listed at position #7 of the November 2018 Top500 list. This system comprises 1088 half-width, 2U CX2570 M4 servers [146], a very dense solution that integrates 2 Intel Xeon Skylake processors and 4 NVIDIA V100

GPUs each, with the latter interconnected through NVLink. This system is optimised for AI applications.

Currently, Fujitsu is working towards the deployment of the Post-K supercomputer [147], which is expected to be a system providing a performance of multiple 100 PFlop/s in 2021. This system is based on custom developed technologies, most notably the A64FX processor [148] and the Tofu Interconnect D (TofuD) network [149], a follow-up of the Tofu1 network used for the K Computer and Tofu2 used for the FX100 systems.

6.4 HPE

HPE is a well-established producer with a broad product portfolio needed for most HPC installations. Not only the compute part now featuring also the ARM platform, but networks from both Intel and Mellanox, Ethernet, storage featuring new approaches like the all flash distributed WekaIO, cooling and power infrastructure supporting both traditional air-cooled installations as well as the direct liquid cooled (DLC) ones. On the x86 platform, HPE supports both Intel and AMD CPUs with an offering ranging from single socket to 64-socket NUMA systems, running a single OS instance.

The high end systems are built on the HPE SGI 8600 platform which benefits from the very high density and relies on the DLC to achieve it. On the other hand this platform can not offer some features that the air cooled platform provides. For example there's a limit to just 4 GPUs per node, Intel CPUs and due to the dense form factor also the size of local storage and I/O connectivity. The air-cooled Apollo platforms offer a broader variety of CPUs, GPUs, FPGAs, and even an NEC vector engine accelerator configuration. The Apollo 6000 represents an air-cooled alternative to the 8600, where the Apollo 6500 focuses on many GPUs per node installations (up to eight). Smaller installations with just 4 GPUs per node provide from lower TCO in the Apollo sx40 platform. The Marvel ThunderX2 ARM base CPU is available in the Apollo 70 line, which was developed for the last two years together with major US HPC labs (Sandia, ORNL, LANL and others) and other vendors (Red Hat, Mellanox and others) to provide a complete solution including OS and tuned drivers ready for HPC [121].

This turned into the installation of the biggest ARM-based supercomputer named ASTRA at Sandia National Labs with 2,592 compute nodes, of which each is 28-core, dual-socket and has a theoretical peak of more than 2.3 PFlop/s. On the side of high-end systems, a next generation of the HPE SGI 8600 platform will be used to deliver a big, 5000-node system with 24 PFlop/s to HLRS featuring the AMD Rome CPUs.

As for storage, HPE offers their own product line Apollo 4500, further distinguished for object storage (4510), clustered storage (4520) and Hadoop/big data storage (4530) as well as the integration of third parties' solutions like DDN (both Lustre and GPFS based) or the already mentioned WekaIO. On the Apollo storage hardware traditional software solutions like Lustre and CEPH are offered, but HPE also offers complex data management software which is a complete rewrite of the former SGI DMF, now called HPE DMF7. The components of DMF are mostly open source, but the whole product as their integration is licensed by HPE.

For the traditional air-cooled equipment, a new liquid-to-air system is offered. The name is Adaptive Rack Cooling System (ARCS) and HPE claims to have a high efficiency of air cooling. Specifications claim 30% less water per rack and 75% fewer hook-ups than traditional rear heat door exchangers resulting in PUE in range 1.03-1.05 and offering complete free cooling when ASHRAE W3 conditions are met. One ARCS can be plugged in so that it accommodates up to four 42U/48U racks with a total cooling capacity of 150kW (non-redundant) or 110kW (N+1 redundancy).

6.5 Huawei

Huawei seems to focus on providing HPC systems to business customers. Out of 14 systems on the November 2018 edition of the Top500 list, which are manufactured by Huawei, only 2 are installed at academic sites. The fastest system is listed at position #103 and operated by an energy company in China. Like other Huawei systems it is based on commodity servers, in this case, FusionServer 2488H V5 servers with Intel Skylake processors. This server is a dual-socket, 2U rack server, which indicates that the system is not optimised for high compute density.

In parallel to its x86-based server business, Huawei started to develop an ARM-based server processor through its subsidiary HiSilicon and a corresponding TaiShan server line. The most recent processor generation, the Kunpeng 920, was officially announced in January 2019 [150].

6.6 IBM

IBM's footprint in the HPC market is currently largely based on systems built on servers with POWER processors, high-performance storage products based on the IBM Spectrum Scale (also known as GPFS) and tape-based storage products. Until a few years ago, IBM positioned itself mainly as a provider of fully integrated solutions. This approach changed most notably with the establishment of the OpenPOWER Foundation [141], which was founded in 2013 with the goal of establishing an open ecosystem around the POWER processor. This allowed other suppliers to offer POWER-based solutions[9]. Furthermore, GPFS was made available to other suppliers to create their own product offerings.

IBM itself continues to provide end-to-end solutions based on the aforementioned technologies and positions these as offerings that allow realising HPC systems for scientific computing as well as systems optimised for high-performance data analysis and AI.

After Blue Gene/Q IBM stopped to design and build dedicated HPC architectures. The currently offered HPC architectures are largely based on commodity technology components, e.g. the IBM AC922 server comprising 2 IBM POWER9 processors and up to 6 NVIDIA V100 GPUs. The latter is the building block of the supercomputers Summit at ORNL and Sierra at LLNL, which, as of November 2018, are listed on position #1 and #2 of the Top500 list [1], respectively. The nodes of these systems are interconnected by a Mellanox InfiniBand network, i.e. based on a commodity interconnect. IBM focused on allowing for a tight integration of third-party technologies. The GPUs are attached to the POWER9 processors using NVLink, which allows for a significantly

[9] One example is the DAVIDE supercomputer with its POWER8-based nodes delivered by the Italian SME E4.

higher bandwidth between CPU and GPU compared to PCIe used in other nodes and provides hardware support for coherence of host and device memory. Attachment of the network devices is improved through CAPI (Coherent Accelerator Processor Interface). CAPI is a protocol on top of PCIe that allow CPU and attached devices to share the same coherent memory space.

IBM Spectrum Scale continues to be developed with HPC requirements in mind. It is, e.g., used for the Summit system at ORNL, where it provides access to 250 PByte within a single namespace and a target sequential peak read/write bandwidth of 2 TByte/s [142]. To avoid performance degradations in case of disk failures, IBM introduced a distributed software RAID concept call GPFS Native RAID, which is taken to a next level with Mestor [143]. IBM Spectrum Scale, however, has become also a product for cloud environments. It supports different object store APIs like Ceph and Swift. With the support of OpenStack Swift-on-File [144], it has become possible to seamlessly provide access to objects as files and vice versa [145]. This development is interesting for HPC centers for consolidating their HPC and Cloud infrastructures.

6.7 Lenovo

Lenovo actively invests in the HPC market and is listed with 140 systems in the November 2018 edition of the Top500 list, two of them in the top 20. With SuperMUC-NG at the Leibniz-Rechenzentrum (LRZ) in Munich, Lenovo delivered the world's largest non-accelerated, general purpose supercomputer. Lenovo offers x86-based servers with different cooling technologies from standard air-cooled systems, rear-door heat exchanger (RDHX), up to direct-liquid cooled systems operating at a temperature level that allows heat reuse for highly efficient computing centers like LRZ.

Remarkable in the HPC field is Niagara (#59 in the Top500), the HPC system installed by Lenovo for SciNet at the University of Toronto. This is one of the first large installations of an InfiniBand or OmniPath fabric using a Dragonfly topology instead of a standard (fat-)tree.

Lenovo also offers storage solutions based on Lustre and Spectrum Scale (GPFS) and is the only OEM in the market selling IBM's declustered RAID version of Spectrum Scale (GPFS Native RAID) with its "Lenovo Distributed Storage Solution for IBM Spectrum Scale" (DSS-G) product line.

6.8 NEC

NEC did not introduce anything new since the previous version of this Deliverable [5], with SX-Aurora TSUBASA A500-64 remaining its most powerful platform, with 64 Type 10A Vector Engines.

7 Paradigm shifts in HPC technologies

7.1 AI

HPC is no longer the only consumer of supercomputers. In the past years, we have seen HPC and Big Data converging to what is called now HPDA. Since our last report, AI is gaining momentum in a number of axes:

- AI assists regular HPC simulations: we can call it "augmented HPC". This is, for example, the case when a simulation workflow is driven by AI mechanisms, when some algorithms are chosen based on decision criteria coming from neuronal networks or when a physics library is replaced by some AI tools [152].
- AI replaces HPC simulations: in this case, approximate models proposed by AI can help the designer to eliminate uninteresting solutions without using expensive simulations and use his/her computing budget to fine-tune the selected solution. For an example, see [153].
- AI can also assist HPC during the in-situ post processing of numerical simulations, in order to only extract and store pertinent data and by consequence save time and energy. This could also lead to AI based smart computational steering of simulations.
- AI could also support the development of smart resources schedulers, able for example to foster predictive maintenance or to detect running applications by their thermal/performance signature and apply energy policies of the center and thus save energy.
- HPC is used to create high fidelity training AI databases. The output quality of a neural network is strictly related to the quality of its training set. HPC can be used to train networks using a vast amount of input. [154] is an example of this situation.
- HPC machines can accelerate big data and large network learning [155].

This list will continue to grow in the coming years with the advance of machine learning (including but not only deep learning) techniques/frameworks and technologies such as more powerful GPUs or dedicated hardware (e.g. https://ai.intel.com/nervana-nnp/).

7.2 Heterogeneous architectures

The double objectives of limited power consumption and programmability introduce major constraints on the future system designs. To meet those objectives the current solution is to implement modular heterogeneous solutions [156]. In this approach, the user will no more see a supercomputer as a single homogeneous entity but rather as an aggregation of specialized modules sharing common resources such as parallel file systems or visualization nodes on a central network.

As of today, we see the following types of modules:

- Standard CPU module: will be used for legacy codes that can't benefit from specialized hardware including GPUs.
- Accelerated modules: a number of codes have already been ported to this kind of architectures and benefit from the boost offered by those. The accelerator will mainly be GPUs since Intel's manycore architecture died in July 2018.

- FPGA or specialized ASIC: in some cases, a use case requires a specific hardware solution that can either be demonstrated on FPGAs or run in production using an ASIC. That will be the case for programs that don't evolve often (compilation & synthesis time is not an issue). An example is programs that filter streams of data (e.g. video streams, experimental signals processing).
- AI module: Should specialized hardware be required, a module dedicated to AI.
- Visualization module: Visualization often requires large memories along with beefy GPUs. With the increased use of in-situ processing, such a module can be used in conjunction with another one.

Figure 29 illustrates what can be a heterogeneous system. The number and types of modules depend on the funding and the user community using the system.

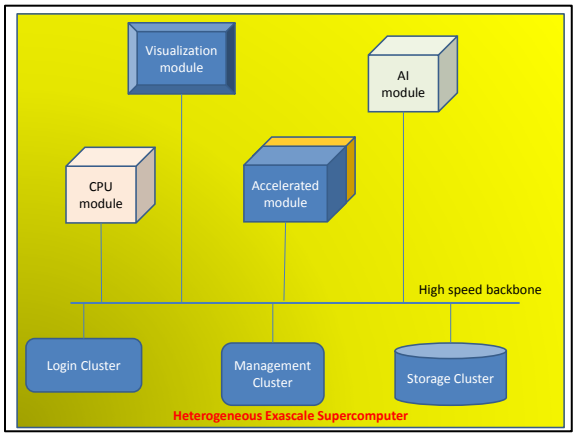

Figure 29: A Potential Exascale system relying on a heterogeneous architecture.

7.3 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. Typical architectural features like lacking processing of instructions, simplified arithmetic or even analogue computations, lacking clock and asynchronous design allow to design neuromorphic devices such that they are very energy-efficient. For some architectures, another benefit is its ability to model neuronal networks faster than biological time.

There is a strong interest in such devices in the context of brain modelling as well as artificial intelligence and machine learning.

In the remaining part of this section we provide an overview of different neuromorphic architectures (in alphabetic order):

BrainScale is an architecture initially developed at University Heidelberg, which is now co-developed within the Human Brain Project [157]. The architecture comprises analogue circuits for modelling neurons, more specifically an Adaptive Exponential Integrate and Fire (AdEx) Model. While the neuron model is largely fixed, the system allows executing these models faster than

biological time. Neuron models running on traditional hardware architectures are typically significantly slower than biological time. The analogue circuits together with digital interconnect logics are integrated into a single waver. Together with other partners, the current BrainScaleS Wafer Modules have been developed, which comprise a silicon wafer, 48 FPGAs and other supporting components. The silicon wafer comprises 384 HICANN (High Input Count Analog Neural Network) chips. 20 BrainScaleS Wafer Modules have been deployed in a single system at University Heidelberg (see Figure 30). A module consumes about 2 kW, i.e. about 5 W per HICANN chip.

Figure 30: BrainScale system at University Heidelberg [157]

Also, Intel's **Loihi** processor implements a spiking neural network process [158] [159]. This device comprises 128 neuromorphic cores and 3 x86 process cores, which are interconnected through an asynchronous network. The design of the neuromorphic cores is digital following a data flow approach, i.e. no instructions are executed. Loihi is explored for use in machine learning problems. For instance, it can be used for LASSO (least absolute shrinkage and selection operator) problems, a regression analysis method. Intel claims the processor to consume tens of mW under load.

SpiNNaker (Spiking Neural Network Architecture) [160] is closer to a standard von-Neumann architecture as it is based on standard ARM cores. The interconnect is, however, brain-inspired with spike delivery to neighbouring neurons based on a fire-and-forget approach. The project was initiated at the University of Manchester and later became part of the Human Brain Project. At Manchester recently, an installation of a large-scale system has been completed. It is composed of 57,600 nodes with 18 cores each, totalling 1,036,800 cores. Power measurements under load indicate an effective power consumption of about 1 W per node [161].

TrueNorth is a reconfigurable processor developed by IBM. Each processor comprises 1 million artificial neurons plus 256 million artificial synapses, which are organised in 4096 neurosynaptic cores [162]. Like SpiNNaker 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 single TrueNorth chip consumes only 70 mW. Research is performed on using this processor in the context of machine learning and artificial intelligence including signal processing (e.g. video tracking, supernova detection), robotics, neural circuit modelling, and optimisation.

8 Conclusion

This second Market and Technology Watch deliverable of PRACE-5IP Work Package 5 gives an updated overview of HPC and Big Data trends, in terms of technologies and with some market and business analysis hints. It is meant to complement the results of the first Technology Watch deliverable D5.1 by highlighting changes that occurred since this deliverable was produced, these changes are numerous due to the fast dynamics of the HPC ecosystem.

The structure of the deliverable is similar to the structure of the previous one. However, in order to provide a better vision of the EU HPC activities, a full chapter describing the related landscape and technology update was added (Chapter 3), while the European Open Science Cloud is presented in the chapter devoted to the cloud (Chapter 2).

The contents of the deliverable are abstracted from a diversity of sources: the Top500 list and publicly available market data and analyses, recent supercomputing conferences (mostly ISC18 and SC18 for this current report), other HPC events and public literature, direct (non-NDA) contacts with vendors and direct participation of WP5 members in a diversity of European projects or initiatives. Beside these projects and meetings, the annual “European HPC Infrastructure Workshop” [163] plays a special role, offering a unique opportunity for WP5 members, experts from the vendor side and from the HPC datacenter facilities management side to collectively learn from problems encountered during datacenter operations. Technical as well as operational aspects related to computing infrastructures are further investigated in Task 5.2, with also corresponding best practices for the design and commissioning of HPC facilities. Best practices regarding prototyping and technology assessment are dealt with in Task 5.3. The combination of these three tasks makes up a consistent and living portfolio of practical documentation for insight and guidance in the area of “HPC Commissioning and Prototyping”.

The work presented in the deliverable will continue in PRACE-6IP Work Package 5 with more focus on building a European view of the worldwide HPC technology and market landscape. This work will be performed with the goal of providing valuable input for the activity of the EuroHPC JU in addition to the technical and site managers involved in PRACE.