E-Infrastructures
H2020-EINFRA-2014-2015

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

PRACE-4IP

PRACE Fourth Implementation Phase Project

Grant Agreement Number: EINFRA-653838

D5.2
Market and Technology Watch Report Year 2. Final summary of results gathered

Final

Version:	1.0
Author(s):	Ioannis Liabotis, GRNET
Date:	21.04.2017

Project and Deliverable Information Sheet

PRACE Project	**Project Ref. №: EINFRA-653838**	
	Project Title: PRACE Fourth Implementation Phase Project	
	Project Web Site: http://www.prace-project.eu	
	Deliverable ID: D5.2	
	Deliverable Nature: Report	
	Dissemination Level: PU *	**Contractual Date of Delivery:** 30 / April / 2017
		Actual Date of Delivery: 30 / April / 2017
	EC Project Officer: Leonardo Flores Añover	

* - The dissemination level 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 2991/844/EC.

Document Control Sheet

Document	**Title: Market and Technology Watch Report Year 2. Final summary of results gathered**	
	ID: D5.2	
	Version: 1.0	**Status:** ***Final***
	Available at: http://www.prace-project.eu	
	Software Tool: Microsoft Word 2010	
	File(s): D5.2.docx	
	Written by:	Ioannis Liabotis, GRNET

Authorship	**Contributors:**	Felip Moll, BSC Oscar Yerpes, BSC Francois Robin, CEA Jean-Philippe Nominé, CEA Guillaume Colin de Verdiere, CEA Carlo Cavazzoni, CINECA Bertrand Cirou, CINES Samuli Saarinen, CSC Susanna Salminen, CSC Dirk Pleiter, FZJ Eric Boyer, GENCI Philippe Segers, GENCI Ioannis Liabotis, GRNET Dimitrios Dellis, GRNET Branislav Jansik, IT4I-VSB Filip Stanek, IT4I-VSB Gert Svensson, KTH Andreas Johansson, LiU Torsten Wilde, LRZ Radek Januszewski, PSNC Norbert Meyer, PSNC Huub Stoffers, SURFsara Walter Lioen, SURFsara Damian Podareanu, SURFsara
	Reviewed by:	Thomas Bönisch, HLRS Florian Berberich, FZJ
	Approved by:	MB/TB

Document Status Sheet

Version	**Date**	**Status**	**Comments**
0.1	15/February/2017	Draft	Initial Draft with TOC and 1st set of contributions
0.2	03/March/2017	Draft	1st almost complete draft with all contributions received before the 02 March included.
0.3	05/March/2017	Draft	Added CRAY contribution, fixed references, acronyms etc.
0.3	07/March/2017	Draft	Added new section in Heterogeneous systems
0.4	08/March/2017	Draft	Minor corrections
0.5	14/March/2017	Draft	Added new section on HPC and Cloud,

			corrections to glossary and conclusions.
0.6	19/March/2017	Draft	Incorporated extra comments and corrections from WP5 partners
1.0	21/April/2017	Final	Addressed internal review comments

Document Keywords

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

Disclaimer

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

Copyright notices

Table of Contents

List of Figures

List of Tables

References and Applicable Documents

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

[2] D5.1: Market and Technology Watch Report Year 1, http://www.prace-ri.eu/IMG/pdf/D5.1_4ip.pdf

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

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

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

[6] CEA and RIKEN Partner for exascale computing using ATM, https://www.hpcwire.com/2017/01/19/cea-riken-partner-arm-exascale/

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

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

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

[10] ETP4HPC, www.etp4hpc.eu

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

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

[13] https://www.top500.org/news/idcs-latest-forecast-for-the-hpc-market-2016-is-looking-good/

[14] http://www.intersect360.com/about/pr.php?id=25

[15] http://www.intersect360.com/industry/downloadsummary.php?id=142

[16] NVIDIA, “NVIDIA Tesla P100 Whitepaper,” 2016.

[17] PRACE Best Practice Guide – Knights Landing

[18] Dongarra, Jack, Report on the Sunway TaihuLight System, June 20, 2016. Online: http://www.netlib.org/utk/people/JackDongarra/PAPERS/sunway-report-2016.pdf

[19] Haohuan Fu et al., The Sunway TaihuLight supercomputer: system and applications. Science China Information Sciences, July 2016. DOI: 10.1007/s11432-016-5588-7.

[20] DDR5, https://www.tomshw.it/ddr5-nel-2020-specifiche-pronte-gia-quest-anno-79378

[21] HBM, AMD http://www.amd.com/en-us/innovations/software-technologies/hbm

[22] Tape Storage Council, http://tapestorage.org/wp-content/uploads/2015/12/TSC-State-of the-Tape-Market-2015.pdf

[23] http://www.netlib.org/utk/people/JackDongarra/PAPERS/sunway-report-2016.pdf

[24] http://www.infinibandta.org/content/pages.php?pg=technology_overview

[25] http://www.slideshare.net/insideHPC/mellanox-announces-hdr-200-gbs-infiniband solutions

[26] OPA User Group meeting at SC16 – University of Colorado Boulder presentation

[27] Discussion with author at SC16 vendor booths – Arista (Koichi Hoydo), Brocade (Pete Moyer), Juniper (Ben Hromyk)

[28] https://www.inphi.com/media-center/press-room/press-releases-and-media-alerts/inphi announces-worldrsquos-first-400gbe-pam4-platform-solution-for-next-generation-cfp8 modules.php

[29] http://www.hpss-collaboration.org/

[30] https://bull.com/sequana/

[31] http://www.cea.fr/english/Pages/News/The-CEA-Tera1000-Bull-Sequana-by-Atos enters-the-TOP500-most-powerful-supercomputers-in-the-world.aspx

[32] http://insidehpc.com/2017/01/atos-delivers-bull-sequana-supercomputer-hartree-centre/

[33] http://insidehpc.com/2016/02/further-expansion-in-the-netherlands-of-bull supercomputer-in-2016/

[34] https://bull.com/mont-blanc-project-selects-caviums-thunderx2-processor-new-arm based-hpc-platform/

[35] https://www.nextplatform.com/2017/01/22/bscs-mont-blanc-3-puts-arm-inside-bull sequana-supers/

[36] http://www.fujitsu.com/global/products/computing/servers/supercomputer/primehpc fx100/index.html

[37] http://www.fujitsu.com/global/about/resources/news/press-releases/2016/1115-01.html

[38] https://www.hpcwire.com/2016/09/08/japan-post-k-computer-hits-1-2-year-speed bump/

[39] https://www.csc.fi/-/high-performance-computing-in-the-cloud

[40] *J. Zhang, X. Lu, and D. K. Panda, "Performance Characterization of Hypervisor-and Container-Based Virtualization for HPC on SR-IOV Enabled InfiniBand Clusters," 2016, pp. 1777–1784.*

[41] https://www.openstack.org/assets/science/OpenStack-CloudandHPC6x9Booklet-v4-online.pdf

[42] https://www.psc.edu/index.php/bridges-virtual-tour

[43] https://www.chameleoncloud.org

[44] https://www.stackhpc.com/hpc-and-virtualisation.html

[45] https://www.dcache.org/

[46] https://irods.org/

[47] PRACE-4IP deliverable D5.4, "HPC Infrastructure Workshop #7

[48] PRACE-4IP WP5 whitepaper, "Datacentre Infrastructure Monitoring

[49] PRACE-4IP deliverable D5.3, HPC Infrastructure Workshop #6

[50] http://www.slideshare.net/hortonworks/apache-ambari-past-present-future

[51] https://fosdem.org/2017/schedule/event/singularity

[52] https://geopm.github.io/geopm/

[53] https://geopm.github.io/geopm/

[54] http://exascale-projects.eu/EuroExaFinalBrochure_v1.0.pdf

[55] http://www.deep-project.eu/deep-project/EN/Hardware/_node.html

[56] https://www.montblanc-project.eu/arm-based-platforms

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

[58] http://www.etp4hpc.eu/en/euexascale.html

[59] https://ec.europa.eu/programmes/horizon2020/en/news/eight-new-centres-excellence-computing-applications

[60] http://www.compbiomed.eu/

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

[62] http://www.etp4hpc.eu/en/sc16-bof.html

[63] https://atos.net/content/dam/global/documents/investor-presentations/atos-to-acquire-bull-presentation.pdf

[64] https://atos.net/en/2014/press-release/show-to-investors_2014_05_26/pr-2014_05_26_01

[65] http://fortune.com/2016/08/11/hewlett-packard-enterprise-sgi-supercomputer

[66] https://www.emc.com/about/news/press/2016/20160907-01.htm

[67] http://1qbit.com/

[68] http://www.magiqtech.com/

[69] T. Sharp, C. Patterson, S. Furber, "Distributed Configuration of Massively-Parallel Simulation on SpiNNaker Neuromorphic Hardware," International Joint Conference on Neural Networks, 2011 (doi:10.1109/IJCNN.2011.6033346).

[70] Human Brain Project, Neuromorphic Computing Platform (https://www.humanbrainproject.eu/ncp)

[71] J. Schemmel et al., "A wafer-scale neuromorphic hardware system for large-scale neural modeling," ISCAS 2010 Proceedings, 2010 (doi: 10.1109/ISCAS.2010.5536970)

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

[73] http://www.hsafoundation.com/members/

List of Acronyms and Abbreviations

AFA	All Flash Array
aisbl	Association Internationale Sans But Lucratif (legal form of the PRACE-RI)
ALCF	Argonne Leadership Computing Facility
API	Application Programming Interface
ASIC	Application-specific integrated circuit
AVX	Advanced Vector Extensions
AWS	Amazon Web Services
BIOS	basic input/output system
BNNS	Basic neural network subroutines
BXI	Bull eXascale Interconnect
CCI-X	Cache Coherent Interconnect for Accelerators
CCTV	Closed Circuit TV
CNTK	Cognitive Toolkit
CPE	Customer Premises Equipment
CoE	of Excellence (for Computing Applications)
CORAL	Joint Collaboration of Oak Ridge, Argonne, and Lawrence Livermore US HPC s
CPU	Central Processing Unit
cPPP	contractual Public Private Partnership
CSA	Coordination and Support Actions (type of Horizon 2020 project)
CUDA	Compute Unified Device Architecture (NVIDIA)
DARPA	Defense Advanced Research Projects Agency
DOE	Department of Energy
DOI	Digital Object Identifier
DP	Double Precision Floating point (usually in 64-bit)
DRAM	Dynamic random-access memory
DSM	Digital Single Market
EC	European Commission
ECP	Exascale Computing Project
EINFRA	eInfrastructure
ESS	Elastic Storage Server
EU	European Union
EUDAT	European Data initiative
EFlop/s	Exaflop/s Exa (= 10^{18}) Floating point operations (usually in DP) per second, also EF/s
EXDCI	European Extreme Data & Computing Initiative
FETHPC	HPC programme of H2020 Future and Emerging Technologies branch
FP7	7th Framework Programme for Research and Technological Development of the European Union (Research and Innovation funding programme for 2007-2013.)
FFT	Fast Fourier Transformation
FGPA	Field Programmable Gate Array
GB	Giga (= 2^{30} ~ 10^{9}) Bytes (= 8 bits), also GByte
Gb/s	Giga (= 10^{9}) bits per second, also Gbit/s
GB/s	Giga (= 10^{9}) Bytes (= 8 bits) per second, also GByte/s

GEO	Global Energy Optimization
GEOPM	Global Extensible Open Power Management
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
GPGPU	General Purpose GPU
GSS	GPFS Storage Server
GTC	GPU Technology Conference
GUI	Graphical User Interface
H2020	The EU Framework Programme for Research and Innovation 2014-2020
HBP	Human Brain Project
HPC	High Performance Computing; Computing at a high performance level at any given time; often used synonym with Supercomputing
HDD	Hard Disk Drive
HMC	Hybrid Memory Cube
HPAC	High-Performance Analytics and Compute Platform (HPAC).
HPCG	High Performance Conjugate Gradient – a benchmark developed as complement or alternative to HPL
HPDA	High Performance Data Analytics
HPL	High Performance LINPACK benchmark (used for TOP500 ranking)
HPSS	High Performance Storage System
HSM	Hierarchical Storage Management
HTC	High Throughput Computing
HTTP	Hyper Text Transport Protocol
IPCEI	Important Project of Common European Interest
IME	Infinite Memory Engine
IOPS	I/O operations per second
ISC	International Supercomputing Conference; European equivalent to the US based SCxx conference. Held annually in Germany.
JSON	JavaScript Object Notation
KB	Kilo (= 2^{10} ~10^3) Bytes (= 8 bits), also KByte
KNC	Knights Corner, Intel® MIC Xeon® Phi™ processors (first generation)
KNF	Knights Ferry, Intel® MIC Xeon® Phi™ processors (prototype)
KNL	Knights Landing, Intel® MIC Xeon® Phi™ processors (second generation)
KNM	Knights Mill, Intel® MIC Xeon® Phi™ processors (third generation)
KVM	Kernel-based Virtual Machine
LINPACK	Software library for Linear Algebra
LHC	Large Hydron Collider
LLVM	Low Level Virtual Machine
LTFS	Linear Tape File System
LTO	Linear Tape-Open
MB	Management Board (highest decision making body of the project)
MB	Mega (= 2^{20} ~ 10^6) Bytes (= 8 bits), also MByte
MB/s	Mega (= 10^6) Bytes (= 8 bits) per second, also MByte/s
MCDRAM	Multi-Channel DRAM
MFlop/s	Mega (= 10^6) Floating point operations (usually in 64-bit, i.e. DP) per second, also MF/s
MIC	Intel Many Integrated Core Processor Architecture
MPE	Management Processing Element

MPI	Message Passing Interface
NAS	Network Attached Storage
NDA	Non-Disclosure Agreement. Typically signed between vendors and customers working together on products prior to their general availability or announcement.
NSCI	US President Executive Order establishing the National Strategic Computing Initiative
OPA	Omni-Path Architecture
PCP	Pre-Commercial Procurement
PFlop/s	Peta (= 10^{15}) Floating point operations (usually in 64-bit, i.e. DP) per second, also PF/s.
PRACE	Partnership for Advanced Computing in Europe; Project Acronym
RI	Research Infrastructure
RIA	Research and Innovation Action (type of H2020 project)
R_{max}	TOP500 system measured (LINPACK) maximum performance
R_{peak}	TOP500 system theoretical maximum performance
SC	Supercomputing Conference; US equivalent to the European based ISC conference. Held annually in U.S.
SKU	Stock Keeping Unit
SoC	System on a Chip
SRA	Strategic Research Agenda
TB	Technical Board (group of Work Package leaders)
TB	Tera (= 2^{40} ~ 10^{12}) Bytes (= 8 bits), also TByte
TCO	Total Cost of Ownership. Includes recurring costs (e.g. personnel, power, cooling, maintenance) in addition to the purchase cost.
TFlop/s	Tera (= 10^{12}) Floating-point operations (usually in 64-bit, i.e. DP) per second, also TF/s
Tier-0	Denotes the apex of a conceptual pyramid of HPC systems. In this context the Supercomputing Research Infrastructure would host the Tier-0 systems; national or topical HPC centres would constitute Tier-1
TCO	Total Cost of Ownership
TSM	Tape Storage System
WAN	Wide Area Network
XML	Extensible Markup Language

List of Project Partner Acronyms

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

Executive Summary

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

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

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

This deliverable is the second one of PRACE-4IP Work Package 5 Task 1, it corresponds to a periodic annual update on technology and market trends. The deliverable builds on the contents of the first one (D5.1) and highlights updated trends in HPC markets over the course of the last year, including extended information on HPC data technologies and systems and highlighting major paradigm shifts in HPC, that have been observed over the last years. This Task 5.1, “Procurement independent vendor relations and market watch”, corresponds to the first objective of Work Package 5. It is thus the continuation of a well-established effort, using assessment of the HPC market, including storage, based on market surveys, supercomputing conferences, and exchanges with vendors and between experts involved in the work package. Trends and innovations based on the work of prototyping activities in previous PRACE projects are also exploited, as well as the observation of current or new technological R&D projects, such as the PRACE-3IP PCP, the Human Brain Project PCP, FP7 Exascale projects and Horizon 2020 FETHPC1-2014 and follow-ups in future Work Programmes. Finally major paradigm shifts and new trends in HPC market are highlighted.

1 Introduction

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

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

This Work Package builds on the important work performed in previous PRACE Projects [1] and updates the first year’s deliverable D5.1 [2] in terms of technology watch, know-how and best practices for energy-efficient HPC Centre Infrastructures design and operations, and prototyping of HPC systems. It aims at delivering information and guidance useful for decision makers at different levels.

Task 5.1 of PRACE-5IP, “Procurement independent vendor relations and market watch”, corresponds to the first objective of Work Package 5. It is the continuation of a well-established effort, using assessment of the HPC market, including storage technologies relevant to HPC, based on market surveys, TOP500 analyses, supercomputing conferences, and exchanges with vendors and between experts involved in the work package. Trends and innovations based on the work of prototyping activities in previous PRACE projects are also exploited, as well as the observation of current or new technological R&D projects, such as the PRACE-3IP PCP, the Human Brain Project PCP, FP7 Exascale projects and Horizon 2020 FETHPC1-2014 and follow-ups in future Work Programmes. Finally, major paradigm shifts and new trends in HPC market are highlighted.

This is the second deliverable from Task 1 of Work Package 5 of PRACE-4IP. It focusses as the first one on technology and market watch only: this means that some best practice and state-of-the-art aspects which were sometimes intertwined with technology watch in past deliverables are now dealt with in other deliverables or white papers (and tasks) of WP5. D5.2 builds on the contents of D5.1 and highlights updates in HPC technology trends over the course of the last year, provides more insights on relevant HPC storage and data technologies and systems and provides incentives on the major paradigm shifts in HPC technology over the last few years.

This deliverable may contain quite a lot of technical detail on some topics, and is intended for persons actively working in the HPC field. Practitioners should read this document to get an overview of developments on the infrastructure side, and how it may affect planning for future data centres and systems. The series of deliverables on “market and technology watch” has been proven very useful to HPC site managers and administrates, of at least PRACE member countries, who get access to a variety of technical information on the HPC technology trends that are applicable to the development and maintenance of the most important HPC facilities in Europe.

This deliverable is organised in 8 main chapters. In addition to the introduction (Chapter 1) and the conclusions (Chapter 10) it contains:

- Chapter 2: “Worldwide HPC landscape and market overview” first uses TOP500, analysed with a geographical, business topical angle, then proposes some extra considerations from other sources, as well as a brief overview of large HPC initiatives world-wide.
- Chapter 3: “Core technologies and components” is a quick overview of processors, accelerators, memory and storage technologies, interconnect technologies.
- Chapter 4: “Solution and architectures” gives some vendor snapshots, and looks at some trends in storage, cooling and virtualisation and cloud delivery.
- Chapter 5: “Data Storage and management” provides insights on storage solutions as well as services that are relevant to HPC.
- Chapter 6: “Infrastructure or Management tools including energy aware tools” is an overview of various aspects of infrastructure management, that highlights information provided in the relevant workshops and PRACE-4IP deliverables.
- Chapter 7: “System tools and programming environment” gives an overview of HPC system management software and trends in the programming environment of HPC systems.
- Chapter 8: “EU Projects for Exascale and Big Data” scans FP7, H2020 and other projects towards exascale.
- Chapter 9: “Major paradigm shifts in HPC market and technologies” provides information on aspects such as artificial intelligence, quantum and neuromorphic computing as well as the consolidation in HPC market, that are considered relevant for the HPC landscape.

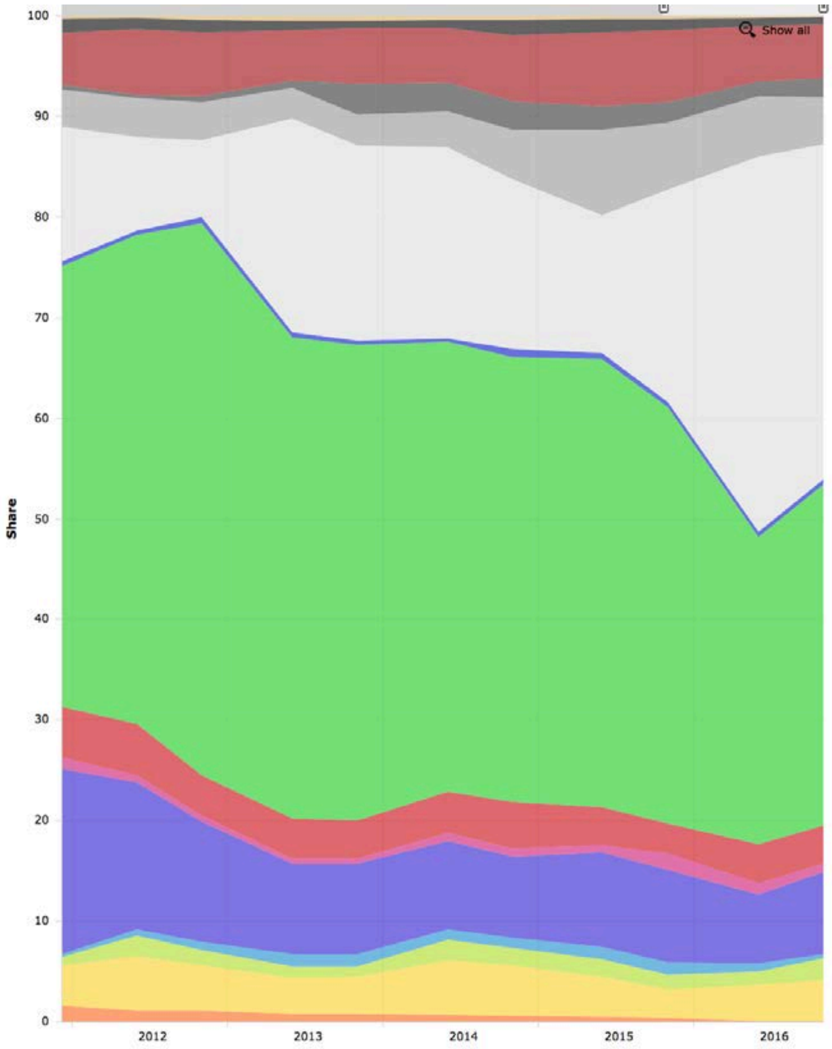
2 Worldwide HPC landscape and market overview

2.1 A quick snapshot of HPC worldwide

The purpose of this section is to present an overview of HPC worldwide, with a special focus on Europe, based on statics derived from the TOP500 [3], as well as information from various sources such as the Supercomputing conference 2016 as well as contacts with specific HPC centres. In D5.2 we focus mostly on qualitative information as a detailed qualitative analysis has been performed in the TOP500 chapter of D5.1 analysing in detail the HPC market as it is reflected via the TOP500, the Green500 [4] and the HPCG benchmark [5] for the last 5 years.

2.1.1 TOP 500

The following figure is taken directly from the statistics service of the TOP500 website [3]. It illustrates the countries with the highest share in performance in the last 5 years, that is from the TOP500 list of June 2012, to the TOP500 list of November 2016.



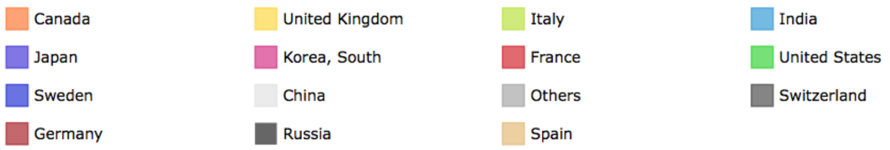


Figure 1 - Countries - Performance share since 2012

The two dominating countries as shown in this figure are the United States and China. What is interesting to observe in this graph is the increase in the share of China since 2013 (Tianhe, Tianhe-2, Sunway TaihuLight) where some very large systems have been installed by Chinese institutions. Japan traditionally also maintains a considerable proportion of the share. From Europe the countries that keep a large share of the HPC capacity are Germany, France, Switzerland, Spain, the UK, Sweden and Italy. All of them are members of PRACE offering Tier-0 (Germany, France, Switzerland, Spain, Italy) or Tier-1 systems (UK, Sweden).

In the November 2016 TOP500 list (Nr 48), China and United States now claim 171 systems each, accounting for two-thirds of the list. China has maintained its ownership of Nr 1 and 2 systems from six months ago: Sunway TaihuLight, at 93 petaflops, and Tianhe-2, at 34 petaflops. Then come: Germany with 32 systems, Japan with 27, France with 20, and the UK with 17.

For reference, last year the following countries were represented: USA with 200 systems, while China had 108, Japan had 37, Germany had 33, and both France and the UK had 18. China is thus clearly keeping and confirming its momentum.

TOP500 systems	USA	China	Japan	Germany	France	UK
Nov. 2015	200	108	37	33	18	18
Nov. 2016	171	171	27	32	20	17

Table 1 - Leading countries in TOP500 shares of number of systems

USA only still have a narrow leadership in aggregate Linpack performance (US with 33.9 percent of the total; China close second with 33.3 percent). Looking at peak performance, China is in the lead with an aggregate of 394 Pflop/s whereas the USA has an aggregate peak performance of 327 Pflop/s. The total performance of all 500 computers on the list is now 672 petaflops, a 60 percent increase from a year ago.

The Top10 is moderately evolving with two new systems in the top ten:

- Nr 5, Cori supercomputer is a Cray XC40 system installed at Berkeley Lab’s National Energy Research Scientific Computing (NERSC), with a Linpack rating of 14.0 petaflops.
- Nr. 6 is the new Oakforest-PACS supercomputer, a Fujitsu PRIMERGY CX1640 M1 cluster, which recorded a Linpack mark of 13.6 petaflops. Oakforest-PACS is up and running at Japan’s Joint for Advanced High Performance Computing (JCAHPC). Both machines owe their computing prowess to the Intel “Knights Landing” Xeon Phi 7250, a 68-core processor that delivers 3 teraflops of peak performance.

Piz Daint, a Cray supercomputer installed at the Swiss National Supercomputing Centre (CSCS), stayed Nr 8 as a result of a 3.5 petaflop upgrade of newly installed NVIDIA P100 Tesla GPUs.

The two most energy-efficient systems in TOP500 are NVIDIA's in-house DGX SATURNV (based on NVIDIA's DGX-1, which houses eight P100 GPUs), 3.3 Pflop/s, and PizDaint. They deliver resp. 9.46 GFlops/W and 7.45 GFlops/W, to be compared with the nominal "exascale" goal of 50 GFlops/W. The DGX SATURNV shows more than a 40 percent jump from the number one system just six months ago on the June 2016 list. The TaihuLight, the number one system, earned a respectable number four position in the November 2016 Green500 with a value of 6.05 gigaflops per Watt.

After years of an ever-increasing share of systems using accelerators/co-processors, in November 2016, we observed a second decrease in a row (following the June 2016 decrease).

The most impressive fact on the TaihuLight system is that there were three different 2016 Gordon Bell Prize finalists exploiting the full 10M core TaihuLight system. The final winner was one of them: "10M-Core Scalable Fully-Implicit Solver for Nonhydrostatic Atmospheric Dynamics.

2.2 Exascale plans China, Japan, USA and Europe

In this short overview, we only present the high-level plans and the underlying technologies with regards to the Exascale plans in China, Japan, USA, and Europe.

2.2.1 Exascale plans China

As of June 2013, China has the world's fastest supercomputer as ranked by the TOP500. In June 2016, the 125 Tflop/s (peak performance) Sunway TaihuLight supercomputer took over the #1 position from Tianhe-2 (55 Pflop/s peak performance). Originally, Tianhe-2 was planned to double its performance by updating and extending the system with Knights Landing, the next generation Intel Xeon Phi processors, however, this was impossible because of the blacklisting by the US government. China managed to double the peak speed of its #1 system by developing its own SW26010 manycore processor.

Current rumors are about deploying Tianhe-2a this year. A presentation at ISC High Performance 2015 from NUDT professor Yutong Lu revealed that next Tianhe system will use general-purpose Matrix2000 GPDSP coprocessors, developed in China, to provide much of its compute power. The Matrix2000 GPDSP is strictly a PCIe-based coprocessor, requiring a host CPU to drive it. Originally, an x86 was envisioned, however, a viable alternative would be ARM64.

The Tianhe-3 exascale system is planned for 2020, the expected power consumption is not specified.

2.2.2 Exascale plans Japan

June-November 2011, Japan had the world's TOP500 #1 supercomputer, named K. This Fujitsu system was based on SPARC64 VIIIfx CPUs and used its own Torus Fusion (Tofu) interconnect. Fujitsu is working on the Post-K Computer. Fujitsu has been collaborating closely with ARM and contributed to the development of the HPC extensions (called SVE) for ARMv8-A, a cutting-edge ISA optimized for a wide range of HPC. As a side note: the Japanese company SoftBank bought ARM for $32 billion in July 2016.

The first exascale Fujitsu system for RIKEN is currently planned for 2020. The power consumption is expected to be 3 – 4 times K's (which is some 12.6 MW).
In January 2017 RIKEN announced a collaboration with France' CEA to prepare for exascale

computing, including efforts to build out the ARM ecosystem - leveraging middleware, applications and workforce training through joint efforts [6].

2.2.3 Exascale plans USA

The current Exascale Computing Project plans in the U.S., foresee that in the period 2023 – 2025 Exascale systems will be in production, together with applications and software that deals with the actual system behavior, where power consumption is in the order of 20 – 30 MW. Currently, three exascale prototype systems are planned in the context of CORAL, a Collaboration of Oak Ridge, Argonne and Livermore.

The US DOE awarded $325 million to build:

- The "Summit" system for Oak Ridge National Laboratory, 150 Pflop/s (2018)
- The "Sierra" system at the Lawrence Livermore National Laboratory, 100 Pflop/s (2017)

Both systems will be using IBM Power9 CPUs, NVIDIA Volta GPUs, NVLink, and a Mellanox EDR InfiniBand interconnect.

The US DOE awarded $200 million to build

- The “Aurora” system for Argonne National Laboratory, 180 Pflop/s (2019)

This will be a Cray Shasta system using the third generation Intel Xeon Phi CPUs, and the second generation Intel Omni-Path interconnect. What is rather special is that the contract is won by Intel and Cray will act as system integrator / sub-contractor.

2.2.4 Exascale plans Europe

Following the February 2012 European Commission (EC) communication [7], recognizing the need for an EU-level policy in HPC addressing the entire HPC ecosystem, all three pillars making up a global HPC chain have been addressed in Horizon 2020 by specific calls in 2014-2015 then 2016-2017 funding periods [8]:

1. next generation of HPC technologies, middleware and systems towards exascale in FET-HPC calls,
2. supercomputing facilities and services incl. support to PRACE via Implementation Phase Projects,
3. excellence in HPC application delivery and use via Centres of Excellence for Computing Applications.

For these periods more than 200M€ have been put in place for the FETHPC and CoE projects. The “contractual Public-Private Partnership on HPC” (cPPP on HPC) [9] organises a framework for the orientations and monitoring of these programmes between the Commission, the European Technology Platform for HPC [10] and the Centres of Excellence. 700 M€ of EC funding in total are provisioned for the HPC cPPP, and the stakeholders are now preparing recommendations for the EC regarding the last period 2018-2020 of Horizon 2020. This includes a vision of pre-exascale HPC systems and solutions integration paths on top of past, current and future H2020 HPC R&D efforts. No specific leading technology/processor track has been selected so far.

A general trend and a framework for analysis is the intertwining of HPC with a growing number of industrial applications and scientific and societal domains. This places HPC as one of the key contributors to the Digital Single Market (DSM) strategy being announced by the EC in April 2016, which actually confirms and widens the scope of the 2012 EC strategy [11]. This includes the notions of a European Science Cloud for scientists to access an underlying rich infrastructure with computing, data storage, access and processing as well as networking capabilities.

Meanwhile, in November 2015, an initiative led by a few Member States (Luxembourg, France, Italy, Spain) has also been announced, so-called IPCEI HPC-BDA, an Important Project of Common European Interest mixing HPC and Big Data objectives [12]. This is understood as an action with strong industrial structuration, from volunteering countries wanting to optimise and align different aspects of a large European HPC and Big Data initiative with the objectives of the DSM.

2.3 Business analysis

IDC bi-annual review of the latest trends in the HPC market [13] was presented at SC16. Other sources (e.g. Intersect360 [14][15]) deliver the same kind of vision for the global trends, if not for the exact figures.

The overall HPC market is doing well, a regular sustained several percentage annual growth is predicted by all:

- Servers are still the largest segment, but storage is a faster growing segment (related to big data / data analytics increasing role and demand).
- The very high-end segment is more irregular (mostly driven by cycles of public investments that may face austerity in some regions).
- Industrial and commercial HPC is a strong growth driver (stronger than the academic sector).
- Cloud computing is still a relatively small segment, although a raising sustained high interest is observed.
- High-performance data analytics is confirmed as fast growing and expanding market segment; AI and cognitive high-performance computing needs are becoming more and more apparent and should have an increasing role in the future.

The trends are general but the situation is not fully homogenous in all regions. The Chinese market in particular has been comparatively moving faster. Also, a growing role of Chinese integrators in the HPC and HPDA related markets is observed (Chinese players grew from almost nothing between 5 to 10 years ago). Some of them, before starting an offer for HPC and HPDA, were already large companies in other technology market segments (mobile, network, PC, etc.). For most of them the main focus is on hyperscale/server/cloud markets motivated by domestic cloud players. They have demonstrated to be able to develop and build custom systems as well as general-purpose servers. They are able to integrate their own silicon as well as other state of the art HPC silicon. Again, they have a complete offer: server, network and storage, and they are already strong in a similar market (server for hyperscaler data s). Finally, they have demonstrated to be able to leverage technologies from other market segments. To this trend an important contribution has come from the acquisition of the x86 server business from IBM by Lenovo. Nevertheless others are now starting to enter the western markets with high-quality enterprise products, like Huawei and Inspur.

What was said above can stand as well for ‘hyperscalers’ in general, not only in or from China. ‘Verticalisation’ is observed under other aspects, from historical component suppliers. Intel various acquisitions (such as ALTERA for FPGAs) and moves (OmniPath for interconnects) are the most salient example affecting HPC: Intel is more and more in a position to become integrator and prime contractor, providing whole solutions from its components, blurring the distinction with the integrators that made up their business - their historical partners and customers. It can be noted IBM used to be in such a position of vertical provider, but moved towards a strong logic of global services and discontinued some of their technological tracks, while becoming less focussed on integration (OpenPower main purpose is about ecosystem development with a possible diversity of integrators involved).

The question of integration added-value from a technological and business standpoint will likely be reshuffled or affected in the future. In the past solution providers were different than component providers, and mostly led the customer and end-user interaction when it came to designing and integrating HPC systems. Today, the borders are more blurred, component suppliers start to have an offer for integrated solutions (such as NVIDIA DGX-1, Intel White Box), and have a growing role in whole system design, not to mention a major influence on pricing (with strong impact on the global solution cost, even when they are not prime integrators). Tomorrow pure integrators might have to find their role, between hyperscalers able to short-cut the supply and technology value chain, and component suppliers expanding and scaling up their business.

3 Core technology and components

3.1 Processors

Following the first market watch deliverable, things haven't changed that much in the processor world. Multicore architectures and further miniaturization (hello 10nm) will be the main course of action for this year.

3.1.1 AMD

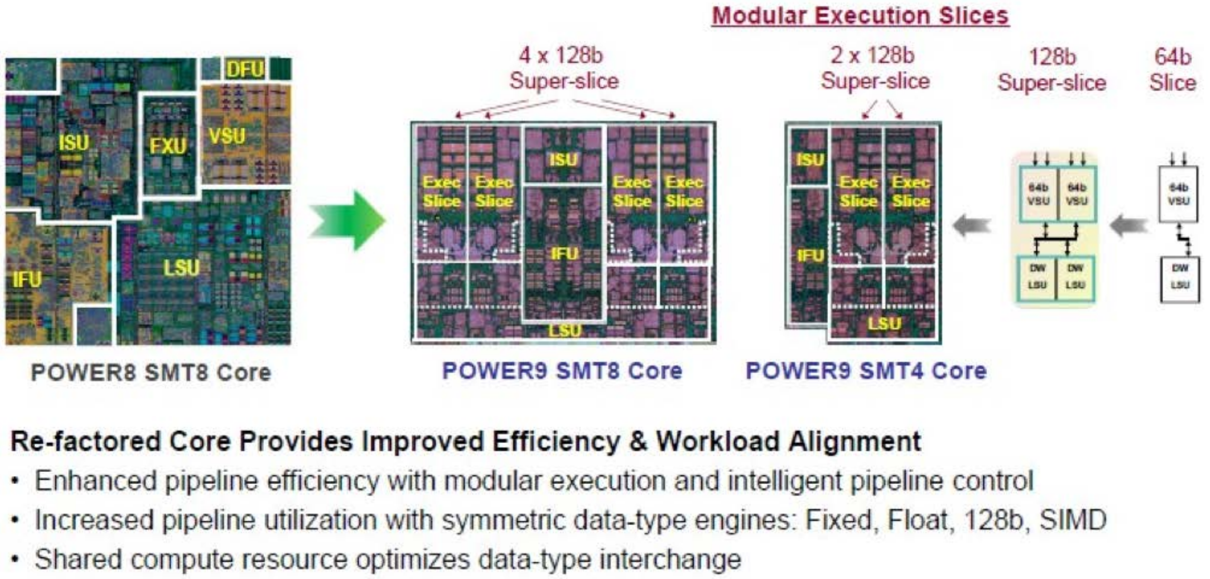
AMD's first ZEN processor is finally here. Ryzen is the official brand name for desktop solutions, and Naples the server edition, which will be the first Xeon competitor that Intel has encountered in several years. With 32 cores and 64 threads, eight DDR4 channels, 64MB of L3 cache and increased queues for both integer and floating point operations, Naples is set to compete with new Intel's Xeon Skylake. Naples will also support 128 lanes of PCIe Gen3. Also two possible setups are presented, in 1U and 2U. This new processor is to be expected in the Q2 2017.

3.1.2 ARM

ARM is following a dubious path. Many companies have tried, without success, to walk into the data with an innovative solution built upon ARM processors, like Calxeda, Samsung or Broadcom to name a few. Others have not given up yet, even with the mediocre adoption of their products: AMD, Applied Micro and Cavium. Right now, Qualcomm seems to be the most aggressive in its intent to bring out an ARM server chip that can compete with Intel's Xeon with its 48-core chip, the Centriq 2400. It has announced a joint venture with Huaxintong Semiconductor Technology to deploy an ARM chip for the Chinese market. On the other hand, Fujitsu will present its Post-K supercomputer based on the ARMv8 SoC in one or two years, which aims for the Petaflop. Still, it's a cloudy future for the ARM vendors, but only time will tell.

3.1.3 IBM

IBM will soon present the last iteration of its own architecture PowerPC, the Power9, with 24 cores built in 14nm FinFET and with 120MB L3 cache. IBM is aiming for optimized technical/HPC workloads, hyperscale, analytics and machine learning applications.



Re-factored Core Provides Improved Efficiency & Workload Alignment

- Enhanced pipeline efficiency with modular execution and intelligent pipeline control
- Increased pipeline utilization with symmetric data-type engines: Fixed, Float, 128b, SIMD
- Shared compute resource optimizes data-type interchange

Figure 2 - The Power9 architecture

The new cores will come in two flavours; a “scale-up” design meant for large HPC applications and supercomputers with support for four or more CPU sockets and a “scale-out” design with 1-2 sockets. Some of the variants will also support IBM’s Centaur memory buffer technology to improve performance and implement an L4 cache. It will be the first chips to implement Power ISA 3.0, the first version of the ISA to be released since IBM launched its OpenPOWER initiative and the first to support AltiVec 3 instructions. The new cores and their peripheral accelerators will be tied together by its custom on-chip fabric. L3 cache bandwidth is said to be 256GB/s, with support for PCI Express 4.0 and multiple 25Gb/s “Bluelink” ports: the aggregate bandwidth from all these sources is over 7TB/s. There is no official release date, but it should be expected near 2020.

3.1.4 Intel

Intel has been the leader in server processors for a while now, with AMD, ARM and IBM trying to get a piece of the cake. To keep up with the first position, Intel is relying on two ways: the Xeon and Xeon Phi.

Still under NDA, the new Intel Xeon Skylake processor, is expected provide more than the 22 cores and more than the 44 threads of its predecessor, with AVX-512 instructions, delivered in 14nm and combined with Intel OPA (that will reach the 100Gbps mark).

The second row of processors will continue the series initiated with Knights Ferry. More details on this line of processors by Intel is given at section 3.2.2

The future of Xeon Phi will be led by the Intel Knights Mill, which will be focused on deep learning and built in 10nm.

3.2 Highly parallel components/compute engines

3.2.1 NVIDIA Tesla P100

Mid 2016 first products based on NVIDIA's new Pascal architecture targeting the HPC (and machine learning) market became available, namely the NVIDIA Tesla P100 [16]. The architecture had already been announced at GTC 2016 in April 2016. The main features of a Tesla P100 device with a single GP100 GPU are:

- Significant increase of both the single device level parallelism as well as the clock speed resulting in a leap in throughput of operations, including floating-point operations.
- Instead of GDDR5 a new memory technology called HMB2 was introduced, allowing for a significant increase in memory bandwidth.
- NVIDIA introduced a new, proprietary link technology called NVLink to enable significantly faster device-to-device and device-to-host data transfer.
- Enhanced support for managing data transfer between host and device and new support for task pre-emption at instruction-level granularity.

Compared to the K40[1], which was based on the previous Kepler generation, the number of streaming multi-processors (SM) increased from 15 to 56. Although the number of 64-bit FMA pipelines per SM decreased from 64 to 32, the overall number of 64-bit FMA pipelines increased by a factor 1.9. The number of double-precision floating-point operations thus increased from 1920 to 3584 Flop/cycle. Additionally, the (base) clock frequency was increased from 745 to 1328 MHz. The peak performance thus is 4.76 TFlop/s at base clock. The Tesla P100 may be operated with a clock speed of up to 1480 MHz.

The new HMB2 memory technology allows for very high data transfer rates using extremely wide data buses, in this case 1024 bit. As the number of data lines would quickly exhaust the number of available pins, the memory needs to be integrated into the package with the GPU. In this case the HMB2 memory is connected to a silicon interposer. To realise a sufficient amount of memory capacity, several memory dies need to be stacked. Figure 3 illustrates this packaging concept. In total there are 4 HMB2 memory stacks per Tesla P100 providing a capacity of 16 GByte and a nominal bandwidth of 720 GByte/s (compared to 12 GByte and 288 GByte/s in case of K40.

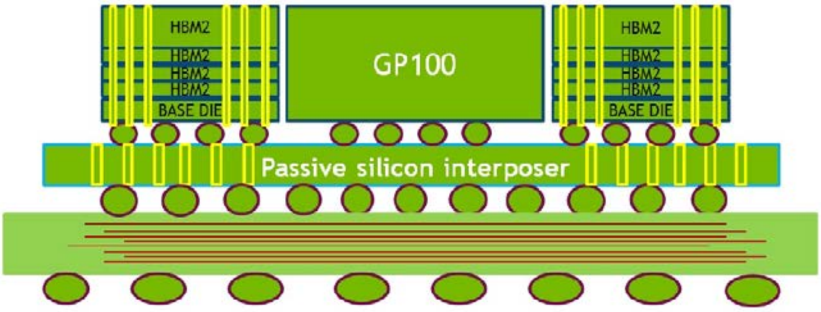


Figure 3 - Cross-section illustrating the packaging of the new Tesla P100 GPU

[1] We use K40 for reference as the K80 is based on 2 GPUs integrated within a single device.

The new NVLink technology allows for a bandwidth of 20 GByte/s per link and direction. Up to 6 links can be connected to a GP100. This is a huge increase compared to a single 16x PCIe GEN3 interface of a K40 with a nominal bandwidth of 15.75 GByte/s per direction.

The enhanced hardware support for managing data transfers strengthens the already earlier introduced Unified Memory technology. This makes on the one hand programming of GPUs easier as the user does not have to care explicitly about data transfers. On the other hand, the new support of page faulting can help to make more efficient use of the available device memory capacity as data transfers can be limited to the really needed data.

3.2.2 Intel Knights Landing

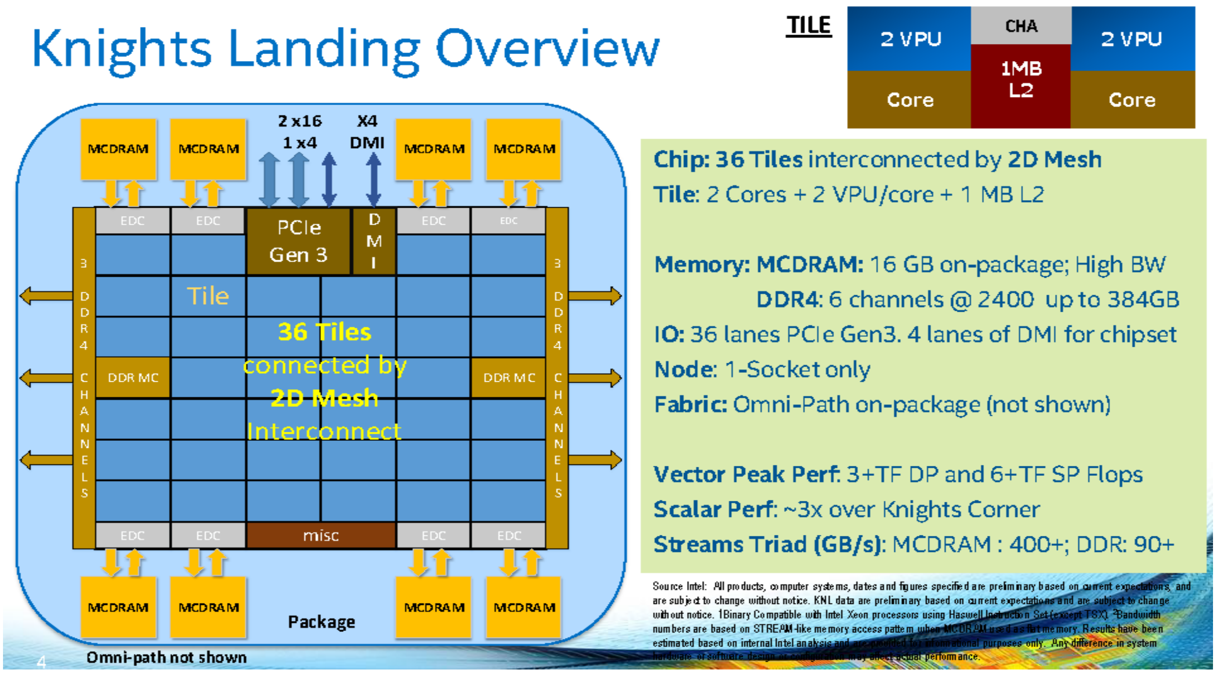
On 20 June 2016, Intel officially launched the Intel Xeon Phi product family x200 based on the second generation Many Integrated Core (MIC) architecture with code name Knights Landing (KNL).

- The first option is a standalone host processor. This option has the main advantage that it can boot and run a full-fledged OS. Another advantage is that it has both: access to the high bandwidth memory and to the much-larger system DDR4, without being slowed by PCIe access. For communication with other KNL nodes, the system I/O fabric is used.
- The second option, and perhaps most appropriate for HPC is the standalone host processor that also has the Intel Omni-Path communication fabric integrated on the package. The latter is denoted by the suffix F in the model number. It provides a 100 Gb/s link across the KNL computing nodes, with arguably lower latencies than InfiniBand EDR.
- The third, similar to the previous incarnation of the Xeon Phi architecture (Knights Corner), is a co-processor card. This card has access to 16 GB of high-bandwidth memory and PCIe access to the host memory.

The Knights Landing processor architecture is composed of up to 72 Silvermont-based cores running at 1.3 – 1.4 GHz. The cores are organized into tiles, each tile comprising two cores, and each core having two AVX-512 vector processing units. Each tile has 1MB of L2 cache, shared by the two cores, for a total of 36 MB of L2 cache across the chip. The tiles are connected in a 2D mesh topology. The cores are 14 nm versions of Silvermont, rather than 22 nm P54C used in Knights Corner. Intel claims that the out-of-order performance is vastly improved, the KNL cores delivering up to 3 times the single-core performance of the Knights Corner cores [17]. The architecture is depicted in Figure 4.

Each of the 72 cores provides out-of-order execution and is multithreaded, supporting 4 SMT threads, similar to the Knights Corner. However, in order to reach peak performance for KNC, one needed to use at least 2 threads/core. In case of KNL, it is claimed that peak performance can be achieved by using 1 thread/core for certain applications. Another advantage of the KNL cores is that they are ISA-compatible with the regular Xeon cores, and can thus run any Xeon application without recompilation.

Another distinct feature of KNL is the 16 GB on-package high-bandwidth memory based on the multi-channel dynamic random access memory (MCDRAM) technology. There are three configuration modes of HBM: Flat mode (same address space, software managed), Cache mode (software transparent memory side cache) and Hybrid mode (part cache, part memory: benefits from both).



Knights Landing Overview

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

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

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

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

Figure 4 - Nights Landing Overview

On 14 November 2016, the 48th list of TOP500 contained 10 systems using Knights Landing platforms.

3.2.3 SW26010

The SW26010 is used in the Sunway TaihuLight supercomputer, which as of June 2016, is the world's #1 supercomputer as ranked by the TOP500. The system uses 40,960 SW26010 CPUs to obtain 93 Pflop/s on the HPL benchmark [18], while the peak performance is 125 Pflop/s.

The SW26010 is a Chinese "homegrown" 260-core manycore processor designed by the National High Performance Integrated Circuit Design in Shanghai.

The SW26010 has four clusters of 64 Compute-Processing Elements (CPEs) which are arranged in an eight-by-eight array. The CPEs support SIMD instructions and are capable of performing eight double-precision floating-point operations per cycle. Each cluster is accompanied by a more conventional general-purpose 64-bit RISC core called the Management Processing Element (MPE) that provides supervisory functions. Each cluster has its own dedicated DDR3 SDRAM controller, and a memory bank with its own address space. The processor runs at a clock speed of 1.45 GHz.

The CPE cores feature 64 kB of scratchpad memory for data and 16 kB for instructions, and communicate via a network on a chip, instead of having a traditional cache hierarchy. The MPE's have a more traditional setup, with 32 kB L1 instruction and data caches and a 256 kB L2 cache. Finally, the on-chip network connects to a single system interconnection interface that connects the chip to the outside world.

Summarizing: each processor is composed of 4 MPEs, 4 CPEs, (a total of 260 cores), 4 Memory Controllers (MC), and a Network on Chip (NoC) connected to the System Interface (SI). Each of the four MPE, CPE, and MC have access to 8 GB of DDR3 memory. A single SW26010 provides a peak performance of over three Tflop/s.

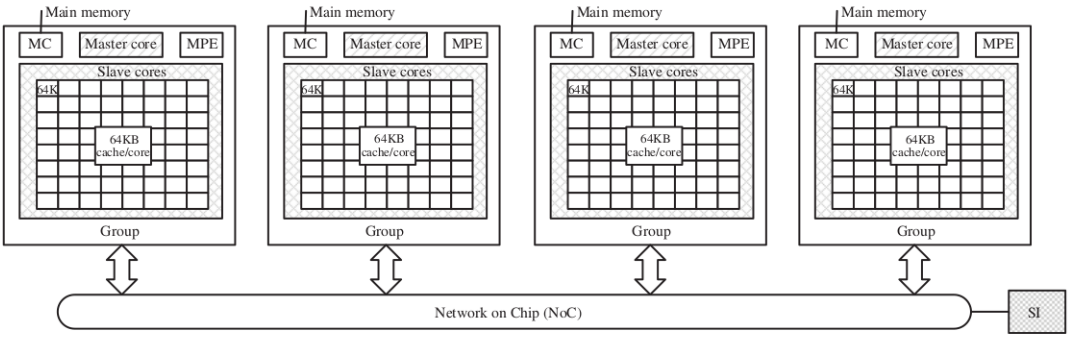


Figure 5: Architecture diagram of the Sunway SW26010 manycore processor chip (source [19])

3.3 Memory and storage technologies

Memory technologies are probably the segment of IT technologies with the highest number of innovations being implemented. Regarding DRAM, new amazing on package high bandwidth solutions are available, while the new DDR5 standard including photonic technologies is being discussed. Beside DRAM, disrupting innovations for non-volatile devices are entering the market, with performance and endurance closing the gap with DRAM technologies. Finally, magnetic media are consolidating their role for long-term storage, with an impressive improvement in bandwidth and density already in the roadmaps. In the following we discuss in more detail DRAM, non-volatile and long-term storage technologies.

3.3.1 DRAM

The gap between the performance of the CPUs and the bandwidth of DRAM continue to widening and the increasing number of cores makes the memory bandwidth limits even more critical. Moreover, the number of memory channels per CPU cannot be increased significantly given the limitation in the number of pins to be used to connect CPUs with memory chips.

Today the DDR4 standard represents the most common solution for system main memory, since DDR5 [20] has not been defined yet. To circumvent the bandwidth limitation, a number of different, non-standard, solutions are applied.

The most innovative solutions are represented by fast DRAM chips, build of 3D stacked memory modules and mounted directly in the chip package, to avoid the limitation represented by the number of pins exiting the socket.

This solution is adopted by last generation GPUs (NVIDIA and AMD), and Intel Xeon Phi.

The main problem of this solution is given by the amount of memory that can fit in the package, which is much lower than the main system memory (e.g. Xeon Phis provide 16Gbyte of on-package memory, for 64 to 72 computing cores available in the same package).

The DRAM memory used by Intel Xeon Phi, is produced by MICRON and derive from the Hybrid Memory Cube. NVIDIA P100 “Pascal” GPUs adopt a similar technology to include memory chips directly on the chip package, called CoWoS® (Chip-on-Wafer-on-Substrate). The memory itself is the HBM2 (High Bandwidth Memory), first introduced by AMD.

Considering that HBM is a memory design from AMD, it is no surprise that the AMD Fiji GPU adopts HBM [21], and future Vega GPU will adopt HBM2 as well.

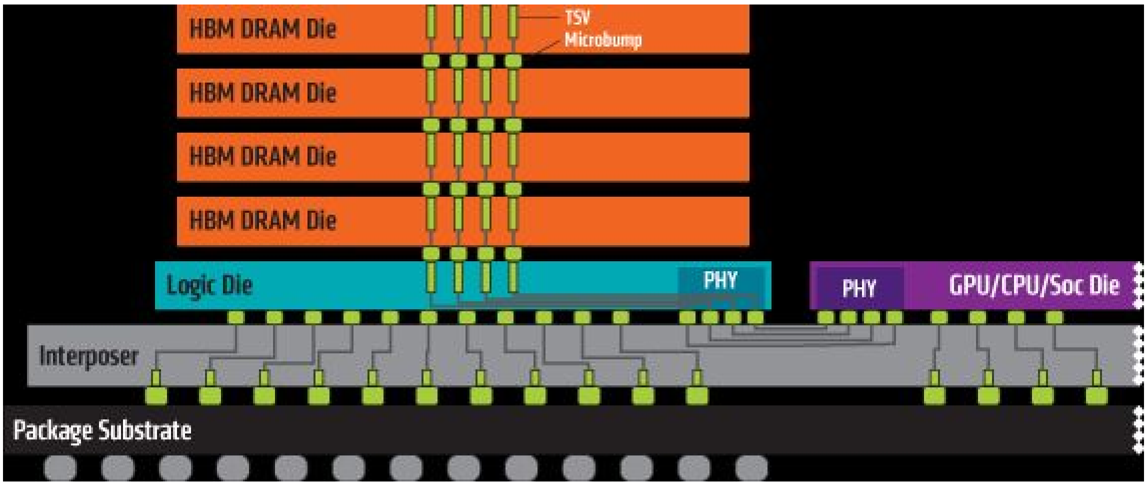


Figure 6 - Schema of High Bandwidth Memory 3D on socket memory integrated in the package with interposer memory logic and processor.

3.3.2 Non-volatile Memory

Concerning non-volatile memory, hard drives continue to scale in term of capacity (e.g. 10Tbyte disk are now available, and 20Tbyte drives are expected before 2020), but not in term of bandwidth and I/O operations per second (IOPS).

This limitation is motivating the adoption of flash systems based, such as SSD (solid-state drive), both with SATA and NVMe interfaces. In particular, the NVMe interface is specifically designed for solid state memory and allows to reach high performance. New innovations for solid state non-volatile memory are expected for 2017. Those will allow making new devices with much higher bandwidth and greater endurance, closing the gap with DRAM chips. Some of these new innovative designs will also allow producing byte addressable non-volatile memory chips, overtaking the limitation of block access.

The performance comparable with DRAM and the byte addressability have the potential to disrupt the HW and SW design of HPC nodes. Here we report a list of most significant non-volatile memory technology innovations.

Micron in partnership with Intel is developing a new solid-state memory called 3D X Point, with the following characteristics:

- Memory Cell based on Material property not on electron storage.
- No transistor is involved in storing data allowing for higher memory density.
- 1,000 times lower latency and exponentially greater endurance than NAND
- 10 times denser than DRAM
- Based on a three-dimensional arrangement of memory cells, allowing the cells to be addressed individually.

Micron reports that 3D XPoint™ technology is an entirely new class of non-volatile memory that can help to turn immense amounts of data into valuable information in real time. With up to 1,000 times lower latency and exponentially greater endurance than NAND, 3D XPoint technology can deliver game-changing performance for big data applications and transactional workloads. Its ability to enable high-speed promises to enable entirely new applications.

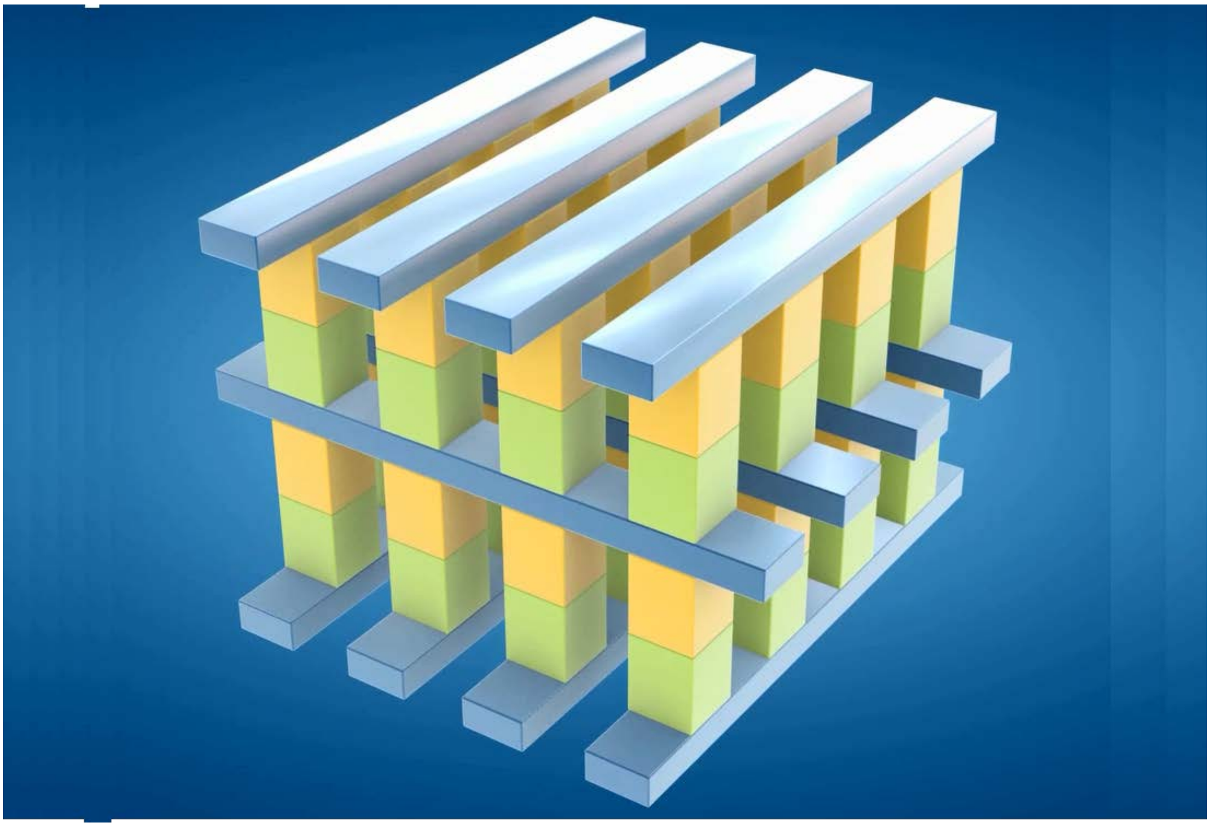
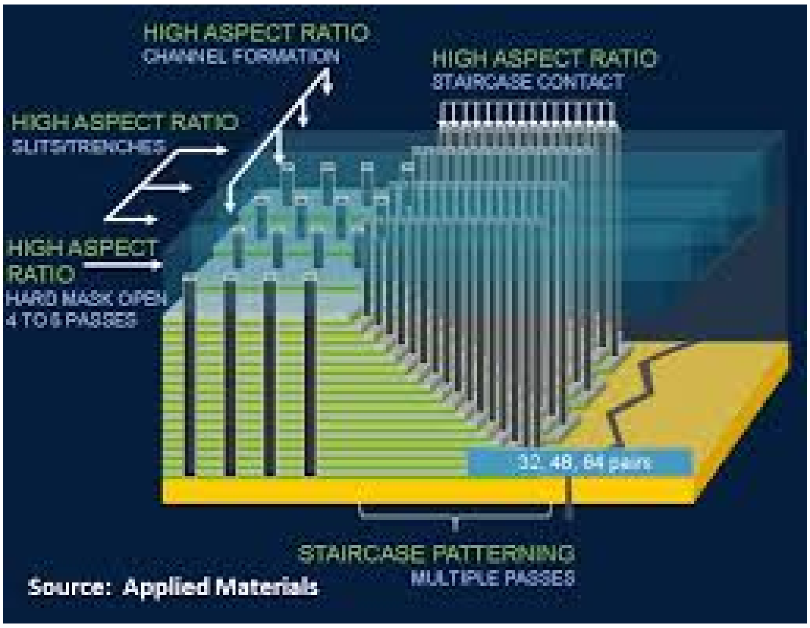


Figure 7 - Schema of 3D Cross Point 3D non-volatile memory design from MICRON: memory cells in Green-Yellow, electrical contact in light blue.

Samsung is developing the V-NAND 3D technology, featuring a unique design that stacks 48 layers on top of each other instead of trying to decrease the cells' pitch size.

Samsung uses Channel Hole Technology (CHT) to enable cells to connect vertically with each other through a cylindrical channel that runs through stacked cells. Samsung V-NAND is virtually immune to cell-to-cell interference.

V-NAND does not need to go through a complex program algorithm to write data and this enables this memory to write data up to two times faster than traditional 2D planar NAND flash memory.



Source: Applied Materials

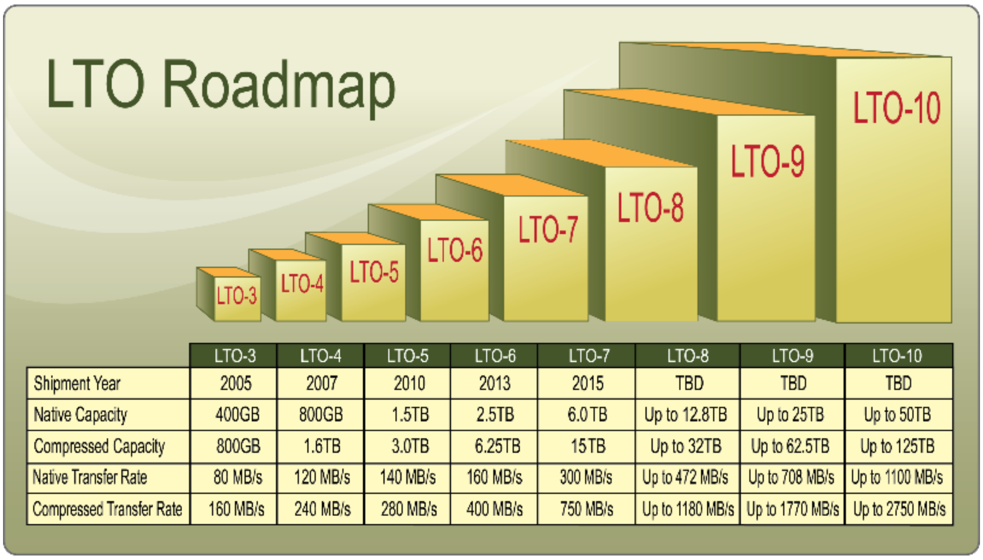
Figure 8 - Schema of V-NAND nonvolatile 3D memory technology developed by Samsung.

Very high-density, high-performance non-volatile memory chips together with NVMe interfaces will allow packing a lot of memory in a small space and could probably compete with traditional DRAM solutions in the design of "fat" nodes for those applications where terabyte of memory are needed in the node. Of particular interest are small form factor standards like M.2.

3.3.3 Tapes

Tape has firmly established its long-term role for effectively managing extreme data growth. LTO and related technologies are becoming a standard, and no real alternatives are available on the market. The main players providing Tape technologies are IBM, Spectra Logic, HPE, DELL/EMC, Oracle.

Steady developments have made tape technology the most reliable storage medium available, now surpassing HDDs by three orders of magnitude in data reliability. As a result, tape is well positioned to effectively address many data-intensive industries including cloud, entertainment, the internet, and high performance computing along with data-intensive applications such as big data, backup, recovery, archive, disaster recovery and compliance.



	LTO-3	LTO-4	LTO-5	LTO-6	LTO-7	LTO-8	LTO-9	LTO-10
Shipment Year	2005	2007	2010	2013	2015	TBD	TBD	TBD
Native Capacity	400GB	800GB	1.5TB	2.5TB	6.0TB	Up to 12.8TB	Up to 25TB	Up to 50TB
Compressed Capacity	800GB	1.6TB	3.0TB	6.25TB	15TB	Up to 32TB	Up to 62.5TB	Up to 125TB
Native Transfer Rate	80 MB/s	120 MB/s	140 MB/s	160 MB/s	300 MB/s	Up to 472 MB/s	Up to 708 MB/s	Up to 1100 MB/s
Compressed Transfer Rate	160 MB/s	240 MB/s	280 MB/s	400 MB/s	750 MB/s	Up to 1180 MB/s	Up to 1770 MB/s	Up to 2750 MB/s

Figure 9 - LTO standard road-map.

Tape technologies have a solid roadmap, with a growing capacity per mm square already planned and demonstrated for three to four generation (see figure LTO roadmap) [22].

Enterprise tape has reached an unprecedented 10 TB native capacity per cartridge with native data rates reaching 360 MB/sec. Enterprise tape libraries can scale beyond one exabyte as exascale storage solutions have arrived.

Enterprise tape drive manufacturers IBM and Oracle StorageTek have signaled future cartridge capacities far beyond 10 TB with no limitations in sight. Recently Fujifilm announced that in conjunction with IBM a new record in areal data density of 123 billion bits per square inch on linear magnetic particulate tape had been achieved. This density breakthrough equates to a standard LTO cartridge capable of storing up to 220TB of uncompressed data, more than 88 times the storage capacity of the current LTO-6 tape.

Prices are going down as well: LTO storage can be as low as 0.8 cents per gigabyte or $8 per terabyte.

Below is a summary of tape's value proposition:

- Tape drive reliability has surpassed disk drive reliability
- Tape cartridge capacity (native) and data rate growth is on an unprecedented trajectory
- Tape has a much longer media life than any other digital storage medium
- Tape requires significantly less energy consumption than any other digital storage technology
- Tape storage has a much lower acquisition cost and TCO than disk
- Tape's functionality and ease of use is now greatly enhanced with LTFS software

3.4 Interconnect

The last year has been very interesting for HPC customers in the area of High-speed interconnect, because the competition has returned to the market. At least for those who buy a typical HPC cluster with compute nodes based on common X86 architecture.

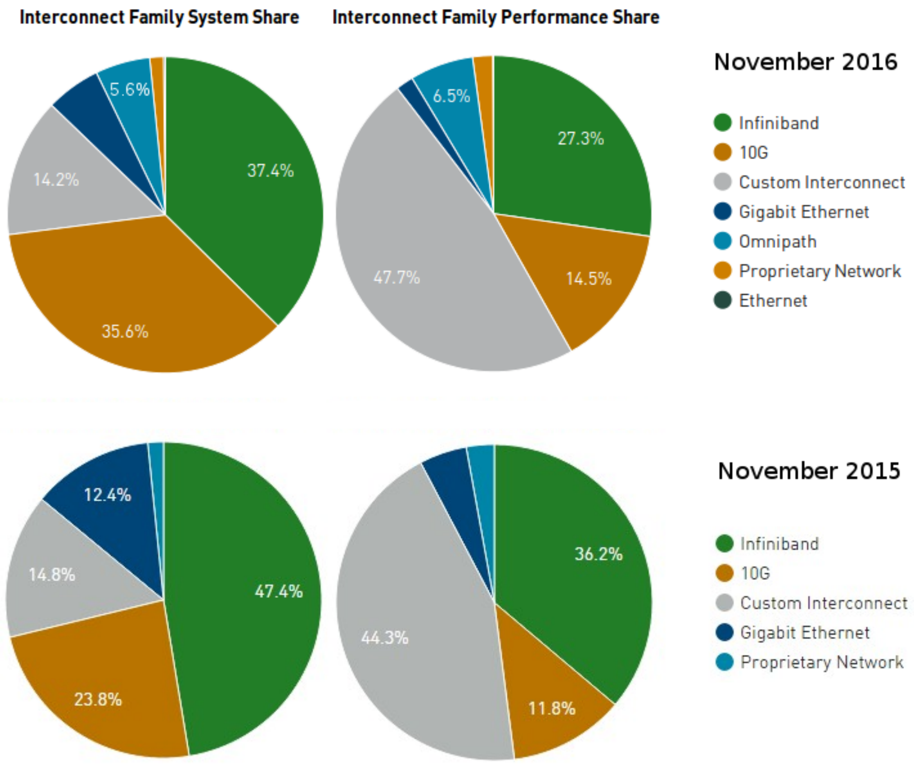


Figure 10 Change of interconnect market (TOP500)

As we can see in the TOP500 [3], Infiniband (driven by Mellanox as the only company) has lost approximately 10% of the TOP500 market. The shift was caused by these factors:

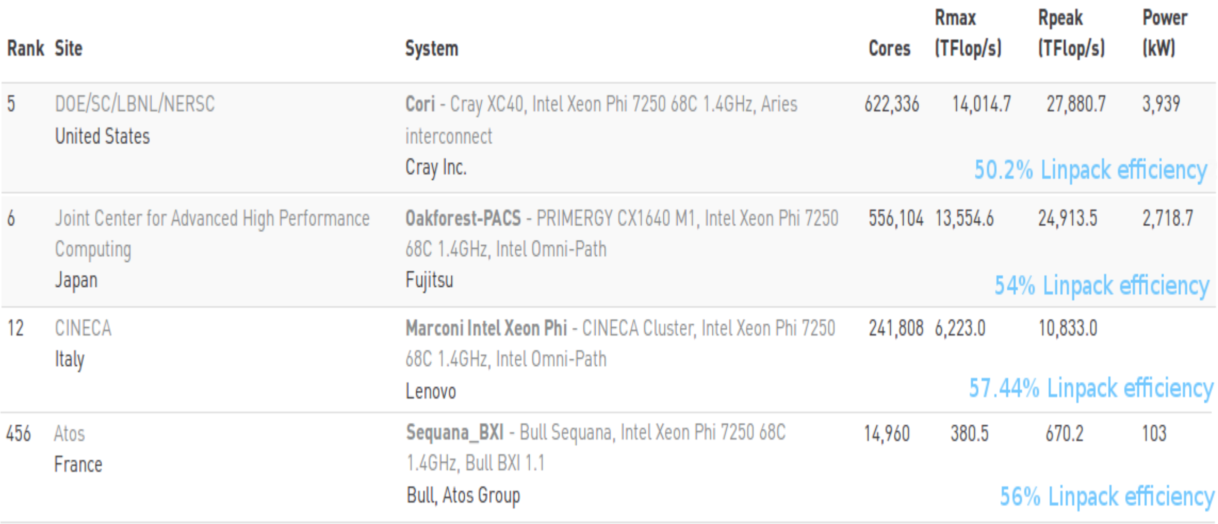
- Intel Omni-Path technology, which started to be generally available in 2016 made it from 0% to 5.6% market share in installed systems base. Omni-Path will be the main factor for interconnect price decrease.
- Fifty-eight new clusters (almost 12% of installed systems) mostly not serving HPC but rather ISPs, SW development, Financial and Web services (e. g. Amazon) equipped with 10G Ethernet were added to TOP500. Main contributors to this trend are China and USA, so it remains a question whether this is driven on the political level of the countries or the national IT developers to get the attraction of their customers.

The number of “proprietary interconnect” systems, which include nowadays only Fujitsu’s TOFU-2 (5 systems) and NUDT’s YH interconnects (2 systems) stays the same as last year.

In the last TOP500 category called “custom interconnects” not much happened in the terms of installed systems with an exception for Cray ARIES systems (five additions), but there are some points to mention:

- The number one of TOP500, the TaihuLight system introduced a new interconnect, which alone affected the performance share of this category by 3%. This is actually a bad marketing for Mellanox, since this interconnect is technically Infiniband (according to [23]). It definitively proofs, that Infiniband as the underlying technology can scale to such extremely large systems as TaihuLight is (40 960 nodes).
- For the first time a production level system with the new Bull Atos BXI interconnect emerged in the TOP500. It is not clear whether Bull will allow other vendors to use this interconnect technology. This might even further increase the level of competition for the HPC users. In the meantime, at least Bull customers have a choice between BXI and Mellanox Infiniband in the Bull Sequana systems. It will be interesting to see some

larger systems with BXI to compare the performance, especially with Bull claiming BXI will offload almost all communication-related work to the network (concept advertised now by Mellanox, too) leaving the main CPU (and accelerators) to focus on the computation.



Rank	Site	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)
5	DOE/SC/LBNL/NERSC United States	**Cori** - Cray XC40, Intel Xeon Phi 7250 68C 1.4GHz, Aries interconnect Cray Inc.	622,336	14,014.7	27,880.7	3,939 50.2% Linpack efficiency
6	Joint Center for Advanced High Performance Computing Japan	**Oakforest-PACS** - PRIMERGY CX1640 M1, Intel Xeon Phi 7250 68C 1.4GHz, Intel Omni-Path Fujitsu	556,104	13,554.6	24,913.5	2,718.7 54% Linpack efficiency
12	CINECA Italy	**Marconi Intel Xeon Phi** - CINECA Cluster, Intel Xeon Phi 7250 68C 1.4GHz, Intel Omni-Path Lenovo	241,808	6,223.0	10,833.0	57.44% Linpack efficiency
456	Atos France	**Sequana_BXI** - Bull Sequana, Intel Xeon Phi 7250 68C 1.4GHz, Bull BXI 1.1 Bull, Atos Group	14,960	380.5	670.2	103 56% Linpack efficiency

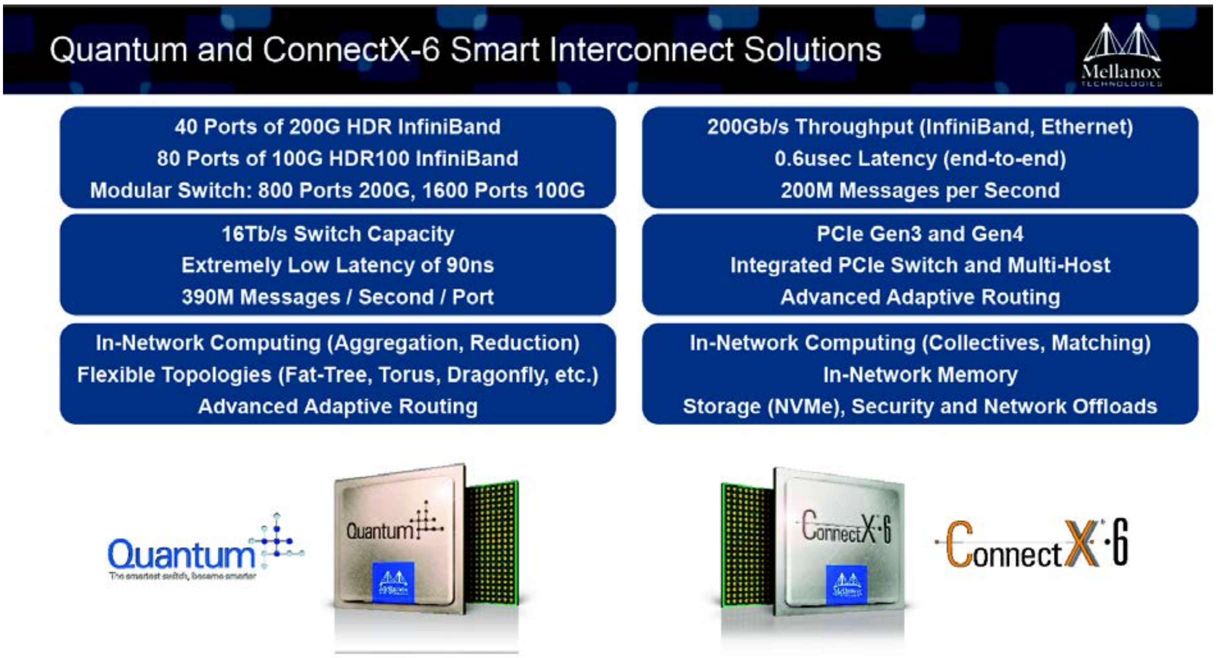
Figure 11 Linpack efficiency for new Intel KNL enabled systems

At this point comparing new KNL based systems (see Figure 11) with Cray Aries, Intel Omni-Path and BXI, we can say that BXI is on par (with 56% of Linpack efficiency). But this is a very rough comparison taking into consideration the various sizes of the systems. Unfortunately, there is not a single KNL system in TOP500 with Infiniband interconnect to compare.

3.4.1 Mellanox Infiniband

Mellanox announced the next generation of Infiniband technology named HDR, which should be generally available in the 4th quarter of 2017. That corresponds to the current version of the technology roadmap [24] and was confirmed at the SC16 Mellanox booth.

Both HCAs and Switch chips will have upgraded specifications (see Figure 12, source [25]).



Quantum and ConnectX-6 Smart Interconnect Solutions

40 Ports of 200G HDR InfiniBand
80 Ports of 100G HDR100 InfiniBand
Modular Switch: 800 Ports 200G, 1600 Ports 100G

200Gb/s Throughput (InfiniBand, Ethernet)
0.6usec Latency (end-to-end)
200M Messages per Second

16Tb/s Switch Capacity
Extremely Low Latency of 90ns
390M Messages / Second / Port

PCIe Gen3 and Gen4
Integrated PCIe Switch and Multi-Host
Advanced Adaptive Routing

In-Network Computing (Aggregation, Reduction)
Flexible Topologies (Fat-Tree, Torus, Dragonfly, etc.)
Advanced Adaptive Routing

In-Network Computing (Collectives, Matching)
In-Network Memory
Storage (NVMe), Security and Network Offloads

Figure 12 Infiniband HDR specifications

Naturally the biggest competitor for the majority of HPC users will be Omni-Path in the next two years.

HDR compared to Omni-Path should offer lower CPU utilization (leaving more CPU time for the computation itself) and promises to have a higher efficiency of the system. This concept is called the “In-Network Computing” by Mellanox and it is introduced already in the current EDR generation of products. It should offload most of the communication work into HCAs and switches.

For those, who will not need 200Gb/s link speed, Mellanox offers to connect two HDR HCAs operating at 100Gb/s to a single switch port operating at 200Gb/s and in this way offering a higher radix than Omni-Path.

The technology offers higher bandwidth, higher message ingest rates and lower latencies. Infiniband in general supports a wider range of compute elements such as IBM/Open POWER CPUs, NVIDIA GPUs (GPUDirect technology) and it is not so tightly integrated with Intel products.

On the other hand, Omni-Path will be integrated into KNL in 2017 and later even into the Xeon CPUs which will reduce latency and probably also reduce TCO. So, until the pricing of the HDR technology is known it’s hard to compare these two rival technologies.

3.4.2 Intel Omni-Path

After one year and 28 production level systems in TOP500 there are some real-life experiences with the Omni-Path interconnect already. The main messages (according to [26]) are:

- IP over Fabric is slower than with Infiniband EDR.
  This seems to be a known issue in Intel which is working on the SW drivers to fix this issue. It needs to be stressed, that this information was valid in at SC16 and might not be relevant at the time of reading this deliverable.
- RDMA/Verbs performance is on par with Infiniband EDR.

> Even in “real life third-party” applications like GPFS, the Omni-Path performs as expected, although after a lot of tuning of BIOS and OS settings, having proper support from the server vendor. There are some differences between Intel BIOS recommendation and the DELL server BIOS used at CU-Boulder.

3.4.3 Cray Aries

Cray with the Aries interconnect is well known to HPC users, but a new view on the interconnect was presented at SC16 during the Machine Learning/Deep Neural Networks and Big Data Analytics session. Cray stressed the importance of Aries interconnect for such workloads, with the focus on:

- Fine-grained RDMA in graph analytics.
- High-performance network well connected to a high-performance memory subsystem for machine learning.

A product called Urika-GX focusing on these challenges was presented, where the Aries interconnect plays a significant role in the overall design.

At this point, this seems like the first attempt to really massively parallelize the Big data analytics/Machine learning and it will be interesting to see other vendors how they will support this market which starts to converge with the traditional HPC.

3.4.4 Ethernet

As already stated in the overview section, most of the TOP500 listed systems with Ethernet interconnect are not really HPC oriented systems. This did not change for the last year as the technology itself hasn’t changed too much. The Ethernet vendors participating at SC16 [27] admitted that their focus is mainly on the scale-out business, telco operators and other customers, who do not need the performance of Infiniband (or similar technology), but require the management and approach of well-known technology and principles offered by Ethernet. The biggest problem of Ethernet in HPC is the much higher latency and higher price per port. Ethernet’s 790ns latency and $1000 per 100GE port compared to 90ns port-to-port latency of EDR and $333 per port can only be compensated with the cost of personnel (training/expertise) to operate an InfiniBand network as well as the need for a complementary infrastructure based on Ethernet technology or services implemented on top of Ethernet-like Software Defined Networks.

The other scenario where Ethernet technology makes sense is WAN technologies, where Ethernet will offer in 2017 an upgrade to 400Gb/s bandwidth [28]. It will be interesting to see, whether the latency will be reduced and thus make this technology viable also in HPC business.

4 Solutions/architectures

4.1 Vendors overview

4.1.1 ATOS

Bull is the ATOS technologies expert brand for hardware and software products, which encompasses, but is not limited to, HPC and Big Data technologies. Bull currently has 20 supercomputers ranked in the TOP500.

Bull’s new generation of Sequana HPC architectures [30] is now commercially available - first installations encompass CEA TERA 1000 in France [31], Hartree in the UK [32] and SurfSARA in the Netherlands [33] CEA aforementioned configuration is part of the second phase of TERA 1000, complementing an already up and running 2.6 Pflop/s bullx system installed in 2016; it is actually a first-of-of-a-kind SEQUANA cell equipped with Bull eXascale Interconnect (BXI), the forerunner of a 25 Pflop/s system to be deployed in 2017. The core feature of BXI is a full hardware-encoded communication management system, which enables compute processors to be fully dedicated to computational tasks while communications are independently managed by BXI.

The fundamental SEQUANA brick is a “cell” comprising two compute cabinets, interconnect switches, redundant power supply units, redundant liquid cooling heat exchangers, distributed management and diskless support. Each cell can contain up to 96 compute blades, i.e. 288 compute nodes, equipped either with conventional processors (such as Intel Xeon processors) or accelerators (e.g. Intel Xeon Phi or NVIDIA GPUs). In each 1U blade, a cold plate with active liquid flow cools all hot components by direct contact – the Sequana compute blades contain no fan.

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

Current Sequana blade configurations are:

- Bull Sequana X1110 blade, with 3 compute nodes, each powered by 2 future generation Intel® Xeon® processors (code-named Broadwell)
- Bull Sequana X1210 blade with 3 compute nodes each powered by an Intel® Xeon Phi™ x200 processor (code-named Knights Landing)
- Bull Sequana X1115 blade with single compute node equipped with 4 NVIDIA Pascal GPUs.

It was recently announced that ATOS is also adopting ARM for HPC, with a special HPC variant of the future ARMv8 “ThunderX2” multicore processors from Cavium. The EU-funded MontBlanc3 project, with BSC, will deploy such Bull’s Sequana HPC hardware with ARM-based ThunderX2 intalled [34][35].

4.1.2 CRAY

Cray continues to provide two product lines of HPC computing systems. Cray XC systems are optimized for very scalable high-end architectures, while Cray CS cluster computers provide higher flexibility. The currently largest systems based on Cray XC is the Trinity system at Los Alamos National Lab (US), with a peak performance of 11 PFlop/s, while the currently biggest Cray CS is a US-government system with more than 6 PFlop/s peak performance, which is currently listed at position #15 of the TOP500 list.

The Cray XC (also known under the codename Cascade) was introduced in 2012. Cray's strategy is to provide a long lifetime for each platform. At the Supercomputing Conference 2016 Cray announced the launch of the Cray XC50 supercomputer with a peak performance of one petaflop in a single cabinet. Cray XC systems are water cooled but there is also XC40-AC as air cooled version available. Cray XC systems have a Dragonfly network topology integrated with Aries network interconnect. Aries reduces communication latency for scale-out applications that rely heavily on MPI interface. Cray XC systems can be equipped with optional DataWarp SSD’s to improve application data performance. New XC50 supports NVIDIA Tesla P100 GPU accelerators and next-gen Intel Xeon and Intel Xeon Phi processors. For system management XC50 also has new image-based management for easy upgrades, less downtime and scalability.

The next generation platform with code name Shasta has already been announced without details being disclosed. Cray has announced that Shasta will merge best technologies from XC line and CS line of cluster supercomputers and is designed to support multiple Intel building blocks, such as future Intel Xeon, Xeon Phi processors and Intel Omni-Path Fabric high speed interconnect technologies. The first implementation of Shasta called “Aurora” is expected to become available in 2018 when the next Argonne Leadership Computing Facility (ALCF) is deployed. Aurora is expected to have a peak performance of 180 PFlop/s.

Cray has also CS400 as cluster supercomputer which have air and liquid cooled designs. The CS product line is highly scalable and based on Intel Xeon E5 processors and supports NVIDIA Tesla GPU’s and Intel Xeon Phi coprocessors.

Cray furthermore continues to provide products optimized for data analytics. The Cray Urika-GX system features Intel® Xeon® E5 v4 processors, up to 22 terabytes of DRAM memory, up to 176 terabytes of local SSD storage capacity, and the high-speed Aries network interconnect. An exclusive feature of the Cray Urika-GX system is the Cray Graph Engine for fast, complex pattern matching and iterative discovery that can run multiple analytics workloads, including Spark, Hadoop and graph, concurrently on a single platform.

4.1.3 FUJITSU

Fujitsu continues to pursue two HPC product lines [37]. Better known is the high-end product line PRIMEHPC based on SPARC processors designed by Fujitsu. The K computer continues to be the flagship installation, which entered the TOP500 in 2011 at position #1 and as of November 2016 was at position #7.

The core of the supercomputer (as of November 2016), the SPARC64 XIfx, 20 nm processor delivers over 1 teraflops peak performance. It has HPC-ACE2 256-bit SIMD extensions (High Performance Computing-Arithmetic Computational Extensions 2). HMC (Hybrid Memory Cube) allows a high memory bandwidth of 480GB/s per node. Tofu Interconnect 2 (Tofu2) is integrated into the SPARC64 XIfx processor and enhances node-to-node communication bandwidth to 12.5 GB/s per link with lower latency. PRIMEHPC is water-cooled. As of November 2016, there were 6 PRIMEHPC systems in the TOP500 (in addition to K), all in Japan and actually ranking down to Nr. 120.

The other product line, PRIMERGY, is based on x86 and Xeon Phi processors from Intel, and has 5 instances in TOP500. In particular, the FUJITSU Server PRIMERGY CX1640 M1 is a modular server node with support for the latest Intel® Xeon Phi™ 7200 product family. It is the base of Oak Forest, a new installation with a performance of about 25 PFlop/s entering rank 6 (just above K) of the latest TOP500.

The University of Tokyo, the University of Tsukuba, and Fujitsu Limited jointly manage the Oakforest-PACS massively parallel cluster-type supercomputer, built by Fujitsu and operated by the Joint for Advanced High Performance Computing (JCAHPC), with a LINPACK performance result of 13.55 petaflops. The system is made up of 8,208 computational nodes using Intel Xeon Phi high-performance processors with Knights Landing architecture that uses many-core processor technology. The nodes are connected by Intel® Omni-Path Architecture [37].

Plans to adopt the ARMv8-A for Post-K had been announced during ISC16 in June last year. In September 2016, Fujitsu's plans have been delayed by up to 12 to 24 months, in order to ensure sufficient processor maturity and volume, keeping the same vision of capable, usable and productive systems with as the overall objective [38].

4.1.4 HPE/SGI

In August 2016 HPE announced the acquisition of the SGI. The acquisition was completed on the 1 November 2016. This merger is indicative of the current trend of system vendors consolidating, and the combined HPE and SGI will have the largest market share of TOP500 systems.

From a TOP500 perspective SGI has been more focused on the upper parts of the list, being a purer HPC and analytics company, with HPE selling many but smaller systems. Looking at it from that angle the merger can be seen as complementary. However, there is also considerable overlap among the systems sold by the companies.

In recent years, the Apollo range has been the HPC focused systems from HPE, building on the earlier SL system (with BL and DL also used by many HPC customers). SGI has a mix of systems with the Rackable series for 1/2U boxes and the ICE systems for larger systems. HPE has always had a wide variety of standard rack mount servers, and it seems that this strategy is still valid. With the addition of the Rackable systems, this will provide an even more varied mix of configurations. There is also some overlap with the HPE Cloudline, which while based on Open Compute and more focused the hyperscaler market, may be price competitive for some HPC workloads.

HPE designed the Apollo 6000 as a purely air cooled system, keeping the water cooling for the Apollo 8000. SGI built the ICE X as a water-cooled system, and later introduced the ICE XA as an air cooled version for data centres where water was not an option. Here the Apollo 6000 is the preferred system for future sales, being more purpose built for air.

The HPE Apollo 8000 was an attempt to break into the high-end of the HPC spectrum with a water-cooled system. It uses a new dry connect design, but is rumoured to have a troubled manufacturing logistics and has not sold in large amounts. This will be discontinued, with the SGI ICE X used as the base for future sales into this market.

4.1.5 HUAWEI

There are four Chinese companies with a significant presence on the TOP500: Huawei, Inspur, Lenovo and Sugon. The dominance of x86 processors on both business and HPC markets allowed these companies to offer solutions designed with cloud and general business markets in mind also as fully functional HPC solutions.

Rising significance of China is visible not only in the business area but also in growing number of machines listed on TOP500 list, both installed in China and manufactured by Chinese companies.

Huawei is active on European and US markets for many years but until recently its activities focused on the telecommunication markets. While HUAWEI is relatively new to the HPC market, the products are mature and reliable as the original target for Huawei servers was mostly cloud operators and telecommunication companies.

Despite having a variety of servers in the portfolio, both blade solutions and standalone servers, two series of products are especially feasible for HPC and are present on the TOP500 list: E9000 and X6800. Both solutions are blade architectures where the second is denser, 8 servers in 4U chassis in X6800 vs 16 servers in 12U in E9000. These solutions are not as dense as the typically

HPC-oriented constructions (e.g. SGI X) but are more flexible in terms of configuration. Typical for HPC, a dual-socket Xeon is featured by a very flexible configuration in terms of memory one can have 16 or even 24 DDR 4 memory slots in a single blade. Due to modularity, both solutions may be equipped with Infiniband or high-speed Ethernet interconnect.

Out of HUAWEI's 16 entries to TOP500 list, the majority of the systems are equipped with Infiniband FDR but six rely on 10 or 100 Gigabit Ethernet as the main network.

While it is possible to install GPUs or Phi accelerators in the Huawei servers, no TOP500 entry was equipped with this kind of hardware. There are also no servers available with standalone versions of Xeon Phi (KNL).

Up to now, Huawei does not seem to have a dedicated HPC solution, rather relying on robust business solutions that can be configured as HPC machines by using Infiniband network and/or installation third-party cooling solutions (e.g. CoolIT).

4.1.6 INSPUR

While the Inspur brand is not very popular outside China, in November's 2016 TOP500 list there are 18 systems based on Inspur's TS10000 platform, all installed in China. The platform is built around dual-socket Intel Xeon servers that can be connected using either Ethernet, Infiniband or Omnipath solutions. Along with the servers, Inspur is providing a cluster management systems: Inspur Tiansuo and inspurCROWN Virtual Cluster suites. The TS10000 platform can be also equipped with integrated storage managed by the same software stack as the rest of the machine. The platform can be equipped with nVIDIA and Intel accelerators.

4.1.7 LENOVO

As is well known, the HPC business of Lenovo started after the acquisition of the x86 business from IBM and in the first two years of activity, they have essentially exploited what was the road map of IBM in the x86 segment. In particular, two product lines were tailored for HPC, the NeXtScale modular system for compute nodes, and the GSS storage system for parallel file systems.

The high-level design goals for the NeXtScale are similar to those associated with previous IBM well known dense platforms (i.e. the iDataPlex), which has been successful in large-scale data and HPC. The intent is to continue to offer configuration flexibility, energy efficiency and cost efficiency in a denser solution that supports a range of network offerings without locking into a particular generation of processor, backplane, or other ever-changing technology.

The Lenovo NeXtScale System is basically composed of three basic blocks:

- An n1200 enclosure, or chassis, responsible for providing power and cooling to the nodes and the nodes expansion.
- A set of servers, or compute nodes, which provide the computational power and I/O connection, either for the production and the management network. The latest server technology is the nx360 M5 which is based on Intel Xeon E5 v3 and v4 CPUs.
- A set of expansions that plug into a node to provide additional capabilities, such as additional non-hotswap HDDs, as well as additional PCI-e slots, typically used to connect GPUs to the nodes.

NeXtScale n1200 enclosure and new NeXtScale nx360 M5 server are designed to optimize density and performance within typical data infrastructure limits. The 6U NeXtScale n1200 enclosure fits in a standard 19-inch rack and up to 12 nx360 M5 servers can be installed into the enclosure.

The Lenovo GSS (GSS21s, GSS22s, GSS24 and GSS26) combines the performance of Lenovo System x servers with IBM Spectrum Scale (formerly known as GPFS) software to offer a high-performance, scalable building-block. The Lenovo GSS allows incremental additions, providing expanded capacity and bandwidth with each additional GSS building-block. The Lenovo GSS features Intel Xeon processors, storage enclosures and drives, software and networking components that allow for a wide choice of technology. Note that IBM as well continues to offer a similar solution but based on Power Processor servers whose market name is ESS.

Beside NeXtScale and GSS, Lenovo offer the xSeries product line (already present in the IBM product line), based on rack mountable 1U-2U form factor, for general-purpose computing and system servers.

More recently Lenovo has re-organized their business units, and HPC becomes part of a larger unit called Data Group, focusing on HPC and hyperscaler business segment.

In 2017, the Lenovo Data Group has announced the introduction of the first product for HPC completely designed by Lenovo since the acquisition of the x86 business unit from IBM. The name of this new server line is Lenovo Stark, whose key characteristic is the modular design.

The Lenovo “Stark” platform will be available, along with the Intel “Sky Lake” processors and will be implemented as a 2U 4 node chassis. The platform will support Apache Pass scalable memory buffer, Omni-Path interconnect and Xeon Phi or other PCIe connected GPUs

4.1.8 SUGON

Similarly to Inspur, Sugon is also a company that is active mainly in China where it is a quickly growing system provider. The product portfolio there are servers designed for general usage, cloud installations and HPC. HPC is represented in high-density blade solutions (TC4600E, TC6600, M-pro Blades) that can be equipped with both Ethernet and Infiniband interconnect. Sugon solutions can be equipped with both Intel and AMD processors that can be supported by nVIDIA accelerators. Using dual-CPU nodes integrated in 10-node chassis one can build an HPC cluster of any size. The largest machine submitted to TOP500 has a rmax of almost 3 Petaflops. Apart from the standard server offer, Sugon has also its own implementation of a Directly Liquid Cooled solution for TC4600E chassis and dedicated HPC platform called Silicon Cube Supercomputer. This machine is equipped with dedicated hierarchical, high-speed network. The machine consists of Hypercubes – sets of servers connected with a non-blocking network. The Hypercubes are connected with each other creating a 3D Torus topology network that ensures scalability and fault tolerance.

4.2 Trends in Cooling of High-Performance Computers

A trend for increased power density becomes more pronounced and is expected to continue for some time. Power densities up to 100kW per rack are becoming common feature present with many HPC vendors. To be able to cool the electronics in high-density computing systems most vendors provide solutions for liquid cooling. Another trend in cooling is the goal to provide a higher temperature of the outgoing liquid. Many computer systems can also operate in higher ambient temperatures than before. Higher temperatures make it possible to use free cooling most of the year also in warmer climates and open the possibility for heat re-use in colder climates. To achieve high temperatures, it is important to capture the heat close to the source and minimize the number of steps in the transfer (for example heat exchangers) to avoid loss of temperature.

A form of liquid cooling is **hybrid cooling** where the electronic components are cooled by air and the air is cooled by liquid close to the source. The liquid-cooling of the air is normally done by heat exchangers in the same rack as the electronics, often as cooled doors or as a more special design. Hybrid cooling has the advantage of being simple to design and the electronics can be the same as in air-cooled versions. For example, cooled doors can be used on most existing racks. However, the heat reuse is often problematic due to high-temperature gradient needed for heat deposition at the air/liquid interface. As a result, the return liquid temperatures are relatively low (10-20 °C) and heat density within the rack is somewhat limited (15-20kW). Additional electric power is needed to move the air through the components, adding to noise and operational costs. Due to the simplicity and low investment costs, we expect this to be used frequently also in the future.

Direct liquid cooling where the electronic components are cooled by the liquid directly is also widely available. Direct liquid cooling enables the high densities mentioned above and the high return liquid temperature (20-50 °C) that allows for heat re-use and for free cooling in most climates. The drawback is higher investment cost for the supercomputer. Higher costs may be compensated by lower costs of the data centre infrastructure as it may be simplified. Operational costs can also be lower than air-cooling, especially in countries with a warm climate. In many cases, direct liquid cooling systems require also additional air-cooling if not all the components in a system are water-cooled.

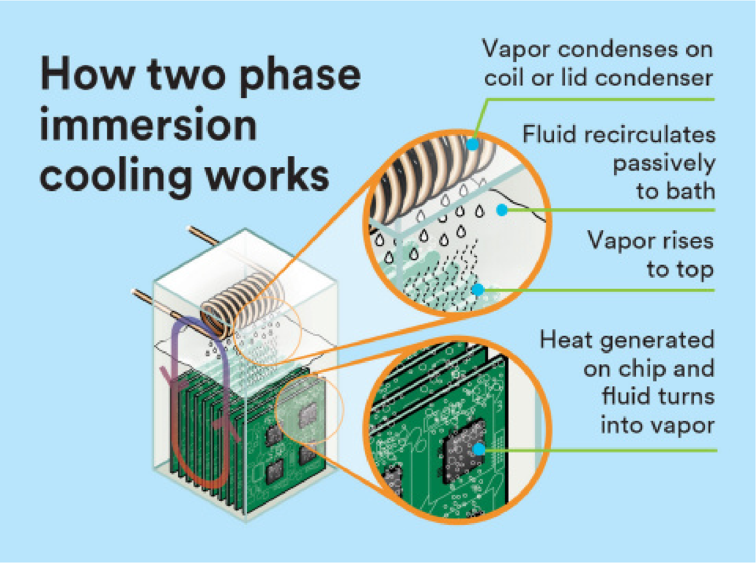

How two phase immersion cooling works
Vapor condenses on coil or lid condenser
Fluid recirculates passively to bath
Vapor rises to top
Heat generated on chip and fluid turns into vapor

Figure 13 - Principles of two-phase immersion cooling

Immersion cooling has been around for some time and starts to gain some momentum. Immersion cooling takes place when we take one or several boards of a system and submerge them in liquid dielectric in a closed or open container. Immersion cooling has the advantage of cooling all the components on the submerged parts equally and the resulting return temperature can be high. In some cases, the boards and the containers are special made and that adds costs to the system. In other cases, we may take ordinary boards, which is more cost efficient.

In general, immersion cooling can be split into two categories, two-phase (or phase change) immersion cooling and single-phase immersion cooling.

Two phase immersion cooling requires closed containers and uses volatile dielectric fluids, such as fluoroketones or perfluorocarbons. Heat generated on chips and other components turns the fluid into vapour (f.x. at 49 °C for heptafluoroisopropyl pentafluoroethyl ketone). The vapour bubbles raise to the top of the container, mixing the fluid and helping the convective heat transfer. Vapour condenses on a condenser, the fluid recirculates passively back to the bath, see Figure 13.

The single phase immersion cooling allows for open containers and uses non-volatile dielectric fluids such as higher alkanes or fatty acids (oils). No phase change takes place, the cooling is due to heat conduction and convection. The cooling dielectric fluid is circulated actively by pumps into the heat exchanger. In both cases, the return temperatures may exceed 50 °C.

Another trend in the cooling of high-performance systems is the **use of volatile cooling liquids** in hybrid cooling or direct liquid cooling. In such cases, the liquid can undergo a phase change (boil) and this can take up a large amount of heat. Temperature loss can be very small and the cooling system may be designed as passive. So far there are no long-term tests of large-scale deployments that let us draw conclusions about this kind of solutions.

Use of dielectric fluids may have a negative environmental impact during manufacturing and disposal. In addition, the usage of volatile dielectric liquids that may change phase at low temperatures (30-80 °C) is posing additional technical challenges for the equipment. Physical properties of this kind of liquids differ significantly from water-based coolants, therefore, active parts (e.g. pumps) of the cooling loop may wear out quickly or behave in a manner that is not tested by the manufacturers.

While immersion cooling and use of volatile cooling liquids provides high return temperatures and lowered electric energy consumption, maintenance can, however, become more difficult and the long-term effects on standard components exposed to the fluid used are unclear. Operational costs and possibly economical benefits are similar to the direct liquid cooling case.

A recent development is the **use of adsorption chillers** that can transform excess heat to cooling. This opens new possibilities to use excess heat from computers for cooling purposes in the computer centre or elsewhere which increases the possibilities for heat re-use in situations where no extra heating is required.

4.3 Trends in virtualisation, cloud delivery

Traditionally High-Performance Computing (HPC) resources comprise bare metal supercomputers and clusters. In IaaS cloud capacity is offered using virtualization technology. On each node a hypervisor runs multiple virtual machines (VMs, or "instances") on virtual operating platforms. A VM can efficiently utilize the central processing unit (CPU) and main memory, but accessing external devices such as disks, graphics processing units (GPUs) and network interfaces may incur significant overhead.

Commercial clouds, such as Amazon web services (AWS) EC2, Microsoft Azure, typically provide fairly standard server nodes, interconnected with Ethernet. The cloud platforms offer different flavours of nodes, optimizing for compute performance, I/O performance, or the amount of memory. The most common interconnect is 10 Gb Ethernet. In the report “High-performance computing in the cloud?” [39]**.** commercial IaaS provider were benchmarked in terms of price competitiveness, and AWS and CSCs Pouta cloud were additionally benchmarked with application benchmarks and micro benchmarks. The conclusions were:

- Existing Ethernet based commercial clouds are not suitable for parallel HPC workloads, since MPI latencies and bandwidth are orders of magnitude worse than on bare metal clusters or supercomputers. Virtualized Infiniband networks, such as the one in Azure Simulation cloud, is expected to perform significantly better [40].
- In-house cloud and containers offer great opportunities for HPC s: Customers with big data workloads and bioinformatics have pioneered the usage of cloud resources. For these workloads the interconnect performance is typically not a major concern. The BioBench2 benchmark also showed that the cloud provides good performance for bioinformatics. Sensitive data, such as patient data, typically cannot be processed in commercial public clouds due to legislation. Even setting up a private cloud that can meet the strict compliance requirements to process such data is not trivial.

Clearly, the interest in HPC clouds is rising. According to Gartner Hype Cycle for Education July 2016, HPC on the Cloud and Computing as a Service had entered the Plateau of Productivity. Many HPC s are developing HPC clouds, some of which are described below. Also commercial clouds are clearly interested in more compute ed workloads. This is evident from, e.g., Microsoft Azure recent offering of HPC cloud with virtualized Infiniband networks, Google cloud offering early access to Intel Skylake nodes, Alibaba cloud launching pilot on Arria 10 FPGAs with Intel and the latest NVIDIA hardware (DGX1 and HGX1) being first publicly available on the cloud. Smaller, HPC oriented cloud providers, e.g Penguin Computing, are also providing Infiniband and Omnipath native high-speed interconnects. The drivers for this are especially complex workflows where large sets of data are handled, which cannot easily be run on bare metal resources; data-intensive workloads in the field of bioinformatics, deep learning on GPU resources, and big data. Although general purpose commercial cloud capacity is still expensive compared to workload optimized on premises installation, it is also economically feasible on cases where absolutely best available performance is required and new technology is adopted immediately upon availability.

The convergence of cloud and HPC technologies is bidirectional. HPC market has been the key driver for performance but hyperscalers are now adopting technologies previously unique to HPC. For example to fully utilize advantages of flash storage and storage class memory a low latency network is required. On the other hand, HPC systems are evolving from traditional Beowulf clusters to cloud approach. One notable example of current trend is Cray’s new system software CLE6.0/SMW8.0 utilizing Openstack and Ansible. Besides the HPC offerings from big cloud providers, public cloud is becoming an important delivery channel for 3rd party HPC services and software vendors via cloud marketplaces.

4.3.1 HPC Cloud Approaches

OpenStack is the de facto stack for providing private and community clouds. Advances in OpenStack support for HPC technologies are strongly linked to their adoption by HPC s. Lately there have been several approaches to improve the HPC performance in the cloud. OpenStack includes modules for integration with Mellanox hardware. This makes it possible to provide InfiniBand HBAs to virtual machines using SR-IOV. This IO virtualization technology provides hardware virtualized network interfaces on the networking hardware itself. For example Monash University has taken this approach for providing HPC capable cloud computing services.

A related technology, PCI pass-through, also enables GPU sharing to virtual machines. PCI pass-through gives virtual machines control of PCI devices and the ability to use them at native performance. This support is now widely available and used by several scientific clouds, including Monash University.

Other approaches include using bare metal provisioning in the clouds where the virtualization overhead is sidestepped completely. This also makes it possible to use hardware with no virtualization support, while benefiting from the self-service cloud model. Univ. of Chicago and TACC Chameleon cloud and the PSC Bridges cloud use this approach.

The popularity of Docker and Kubernetes has also resulted in improved GPGPU and SR-IOV support for Docker. While not widely in use yet, Kubernetes based scientific platforms bring performance benefits over traditional virtualized cloud environments, and better multi-tenant isolation compared to bare metal deployment. The speed and flexibility of container management compared to virtual machine management also benefits HPC workloads.

"The Crossroads of Cloud and HPC: OpenStack for Scientific Research" [41] describes several OpenStack clouds which have been designed for HPC needs. These clouds are summarized here shortly:

Bridges - Bare metal and traditional virtualized workloads
Bridges is PSC's 800 nodes bare metal supercomputer for heterogeneous workloads. Bridges uses OmniPath for connecting all nodes and the distributed 10PB filesystem with 100Gbps speed. HPC applications use RDMA verbs to take advantage of the OP capabilities and OpenStack Ironic is used for bare metal provisioning throughout the system [42]

Chameleon - Bare metal and SR-IOV usage
The Chameleon testbed is a 650 node cloud co-produced by Univ. of Chicago and TACC. The hardware is reconfigurable with Ironic, though a part of the cloud is reserved for KVM virtualization. In addition to x86, Chameleon offers Atom + Arm microservers and GPU (Tesla) accelerators. The testbed has 100Gbit connectivity between sites, 40Gbit between racks and 10Gbit between nodes. Chameleon allows users to map various resources to either advance (utilizing OpenStack's Blazar) or on-demand leases. Chameleon offers e.g. cloud-in-a-cloud (OpenStack), bare metal Docker, KVM hypervisors with SR-IOV over IB, MPI clusters with MVAPICH2 and SR-IOV over IB and also RDMA-Hadoop (with SR-IOV) [43].

MonARCH - GPU and SR-IOV enabled HPC workloads
Monash Advanced Research Computing Hybrid, or MonARCH, is an HPC/HTC cluster which utilizes SR-IOV with 56 GbE in order to achieve the HPC networking requirements. The high-speed network uses layer-2 RoCEv1 in order to enable MPI workloads in the cloud. The newest of the clusters, M3, comprises of 1700 Haswell cores, 16 quad-GPU nodes + one octo-GPU node with NVIDIA K80 dual-GPU accelerators. M3 also offers its users a 1,2PB Lustre parallel file system. Specific OpenStack flavours in M3 enable running e.g. CUDA accelerated HPC workloads [43].

5 Data storage and data management

5.1 Storage Solutions

In this section, we will look at the two dominant storage solution vendors: DDN and Seagate. Both vendors provide complete full-blown scaling solutions for HPC applications. As a rule, the technology provides redundancy at all levels of the implementation; an integrated hardware and software stack and wide connectivity support including Intel Omni-Path, Mellanox EDR and 40 Gb Ethernet. At the bottom level, spinning hard disk drives still provide the bulk data capacity while SSDs and NVMs accelerate the I/O performance, although full flash (SSD) solutions are available as well. At the filesystem level, NFS, Lustre and Spectrum Scale (GPFS) are the dominating technologies. The vendors prefer to pack the products into purpose built "appliances" that provide turnkey, ready to use modules including hardware, software and necessary processing power to interface and integrate the storage to the HPC and other data centre infrastructure.

5.1.1 DDN

DDN structures its storage solutions portfolio into the following categories: Block storage, Flash storage, File storage and Object storage.

The Block storage includes the SFA® product family, currently including SFA12KX, SFA14KX and SFA7700X members. These are hybrid SSD and HDD storage platforms, building blocks for complex storage solutions. Intelligent caching (SFX) and full bandwidth infrastructure deliver performance acceleration from SAS-based SSDs, NVMe and HDDs. Non-blocking internal PCI-e and 12Gb SAS fabrics ensure data access while InifiniBand, Omni-Path, Fibrechannel (FC) and Ethernet provide the connectivity of the module. Sizes from 4U to 84U provides up to 17PB capacity and 60GB/s throughput.

The Flash storage includes the Flashscale AFA and the IME Burst buffer products. Flashscale is the at-scale, enterprise-ready flash storage infrastructure platform. It provides high bandwidth, low latency storage. InifiniBand, Omni-Path and FC provide the connectivity of the module. Sizing from 4U to 40U it provides up to 6PB capacity and 600GB/s throughput. The IME burst buffer is an NVMe based accelerator, that absorbs bulk application data. Built as a standalone commodity server or at rack scale on Flashscale technology, the IME provides a building block for accelerating the I/O performance of the backend systems.

The File storage includes the Scaler family appliances. These "appliances" are complex solution packets, providing all hardware, software and tools to export a file storage services via the NFS, SMB and Lustre filesystem. The Linux and Windows client software and licenses are included. The Scaler appliances may be built all SSD or hybrid, utilizing SSDs for acceleration. InifiniBand, Omni-Path and 10/40/100GbE provide the connectivity of the module. The solution is designed as scalable, the building block being the 5U enclosure. Up to 4 OSS servers, 1PB capacity and 50GB/s per base enclosure, the units were used to build some of the largest single-namespace file systems deployed today.

The Object storage includes the WOS object storage appliance and a suite of hardware and software helper products such as the WOS gateways. The WOS object storage is built from 5U-sized units containing up to 96 drives, up to 960TB of storage. The unit appliance also contains the object storage node, powered by an Intel CPU and 128GB RAM. The unit stores up to 8 billion objects and provides a 10GbE interface. The appliances may be scaled out in clusters of up to 256 nodes, these clusters may be combined into an Exabyte namespace, distributed geographically. The object storage provides NoFS storage, no Linux file I/O or file system in the classical sense. Data replication policy may be defined on a per object basis; self-healing features include data recovery through dispersed data placement. Custom metadata may be attached to the objects. Access mechanisms include multisite ingest, however, the data may be exported as a NAS mount point via NFS or Samba via the WOS Access gateway. Further, the WOS backend may be interfaced to the Scaler products via the GRID-Bridge product for automated data migration to the WOS storage cloud. While the WOS storage may not be the most suitable for direct HPC storage, it is an attractive solution for mass storage and analysis of data obtained in HPC context.

5.1.2 Seagate

Seagate structures its enterprise storage solutions portfolio into the following families of products: ClusterStor, RealStor and OneStor. The AssuredSAN RAID product line of Dot Hill was included into Seagates' portfolio. While ClusterStor is a complex appliance, the RealStor, OneStor and AssuredSAN RAID are more traditional block storage arrays targeted at different market domains.

The ClusterStor family products are scale-out solutions providing Lustre parallel file system for HPC. The ClusterStor systems are complex solution modules, providing all hardware, software and tools to export a file storage services via the Lustre and SpectrumScale filesystem. The Intel Enterprise Edition for Lustre is based on Lustre 2.7 and supports up to 18 meta data file servers per files ystem. The N-product lines (L300N, G300N) are accelerated using NVM based Nytro accelerator. InifiniBand, Omni-Path and 40/100GbE provide the connectivity of the module. The solution is designed as scalable, in principle infinite scalability is claimed. Largest deployments achieve about 1.7TB/s throughput and over 80PB capacity.

The RealStor family products include Hybrid and all-flash block storage arrays. Packaged as 2U enclosures, the arrays provide real-time tiering, thin provisioning, SSD read cache, snapshots and replication. FC, iSCSI and SAS connections are used. 7GB/s throughput and 46TB capacity can be achieved within the enclosure. The all-flash device achieves up to 400kIOPS at 1ms latency.

The OneStor is a modular, standards-based block storage array designed for OEMs. Packaged as 2U to 5U enclosures, OneStor may hold up to 84 SAS and SATA HDDs or SSDs. This allows for over 500TB and 14.4GB/s in a dual controller configuration, using 6TB HDDs. A host/expansion interface is a novel dual 6Gb/s SAS I/O, offering longer-active cables, universal ports, self-configuration and standardized zoning. The modular design allows for combining four such enclosures, achieving a setup up containing up to 336 drives. The OneStor embedded storage incorporates Intel based server modules that insert directly into the enclosure, consolidating storage, processing, memory and I/O channels into a compact device.

AssuredSAN RAID are SAN storage arrays by Dot Hill manufacturer. The family includes the 3000, 4000, 5000 and 6000 series. The 2U to 4U enclosures provide FC, SAS and iSCSI connectivity, real time tiering, read cache and thin provisioning. The 6000 series 4U enclosure targeted at HPC market holds 40GB SSD cache and delivers 12GB/s throughput and up to 448TB of capacity. The enclosure can be expanded with expansion units, adding up to 248 additional drives and 1.9PB capacity. The units may be configured via a web GUI or command line interface via SSH.

5.2 Off-line storage

This section covers storage that is not directly connected to compute clusters as a target for reads and writes of data used in batch jobs. In some cases the data may be used for compute jobs in a more indirect fashion, with two examples being tiered storage systems or data staging in grid environments.

5.2.1 Tape storage

In many ways off-line storage describes any storage medium that is non-volatile and whose data cannot be accessed by the computer once removed. A good example of off-line storage is a tape storage. Tape has been around since the early days of computing. At first used as primary storage, tape is now being used for backups, archiving and less frequently accessed data in tiered storage systems. Access to data is sequential, so while the transfer bandwidth is high the latency is also very high. One of the advantages of the medium is that it can easily be removed from the storage system and be stored in an off-site facility for physical separation.

Much like the hard disk market, the tape storage market has consolidated in recent years. IBM is now the main vendor driving the LTO standard, supplying drives to all tape library vendors using LTO. With the current media pricing LTO-6 is usually cheapest per TB, even though it is quite low density, unless a site is space constrained.

Spectra Logic has traditionally been using IBM drives in their libraries, but started selling Oracle Storagetek T10k as an option for the TFinity library in April 2016.

HPE has traditionally resold Quantum libraries with HP drives, but have now switched over to reselling Spectra Logic libraries with IBM or Oracle Storagetek drives for the high-end.

IBM and Oracle Storagetek, of course, offer their own drives in their libraries, but Storagetek will also sell IBM LTO drives.

Drives/Library	*IBM*	*Oracle*	*Quantum*	*Spectra Logic*
LTO	Yes	Yes	Yes	Yes
3592	Yes	No	No	Yes
T10k	No	Yes	No	Yes

Figure 14 - Vendors and types of drives/libraries

5.2.2 Tiered storage

Hierarchical storage managers (HSM) takes a normal file system and adds different tiers of storage with differing performance characteristics. Most often this is done to keep the most used data on fast low latency storage, and have a larger pool of slower (and presumably cheaper) storage for less often used data. The HSM software handles the migration of data between the tiers and needs to be tightly integrated into the file system.

IBM has two major HSM implementations, one for the Spectrum Scale (previously GPFS) file system using Spectrum Protect (previously TSM) as a storage backend; and High-Performance Storage System (HPSS) that can both be a standalone file system or a Spectrum Scale backend.

HPSS [29] is built for large distributed storage installations, and targets many petabytes of data. One issue with storing large amounts of small files on tape is handled by HPSS by aggregating small files when doing migrations to HPSS tape.

Lustre has historically not had good HSM support, but with the introduction of the Robinhood based HSM support in Lustre 2.5 this has been amended thanks to work done by CEA. Robinhood stores metadata in a database and can act as a policy engine for Lustre and do rule based migrations to other storage tiers.

Spectra Logic has traditionally been a pure tape company, but have recently also produced disk-based products. Their **Black Pearl** products are disk based, but also acts as an S3 based gateway to either tape libraries or their (also S3 based) **Arctic Blue** archival disk subsystem. Writes to underlying storage are size triggered and read caching is done.

5.2.3 Object storage

The term “Object Storage” has become popular as the name for non-POSIX storage, often accessed through an HTTP-like interface. File systems can be layered on top of this to enable accessing the data in a more traditional way, creating what can be seen as a kind of tiered storage system.

In the grid computing world, **dCache** [45] is a common storage solution, coming out of the high-energy physics community and used for storage of data from the LHC experiment at CERN. It is a highly-distributed storage system and, as the name implies, handles local caching of data fetched from remote sites. Data may also be replicated on multiple sites for redundancy, with a global namespace. For accessing tape storage dCache has support for optimizing the order in which files are read to minimize seeking. Apart from its bespoke access protocol there are also multiple other access methods (called doors in dCache) available, such as NFS and GridFTP.

The **Integrated Rule-Oriented Data System (iRODS)** [46] is not a pure storage system, but sells itself as a data management system. It provides a single namespace and metadata handling for underlying storage, but the storage can be spread over multiple storage systems. Apart from the storage functionality it can also be used to implement workflow automation with action rules defined that are triggered by storage events. iRODS is used by the EUDAT project and a number of life sciences sites, for example.

Scality has a software-based solution that works with storage hardware from multiple vendors. Supported configurations are available from HPE as well. The RING software manages the object placement and uses replication and erasure coding to provide high availability. S3, OpenStack, SMB and NFS interfaces are provided on top of the native object storage. CEA is investigating its use as a Lustre storage tier.

5.3 Data services including data analytics

5.3.1 Introduction to Data Analytics

Nowadays, humans and machines, including PCs, tablets, mobiles and HPC clusters, are connected to the global Internet network that carries huge volumes of information. This network allows more efficient communication enabling the interconnection of new data sources including sensors, IoT devices, weather observation stations, CCTV cameras, etc.

Statistics related to current data sources for IT systems are presented in the picture below. They clearly show that the Exabyte size of the data produced and processed around the World was already exceeded several years ago. Analysts claim that most of the data are in fact unstructured information, whose processing is hard or impossible using classic approaches.

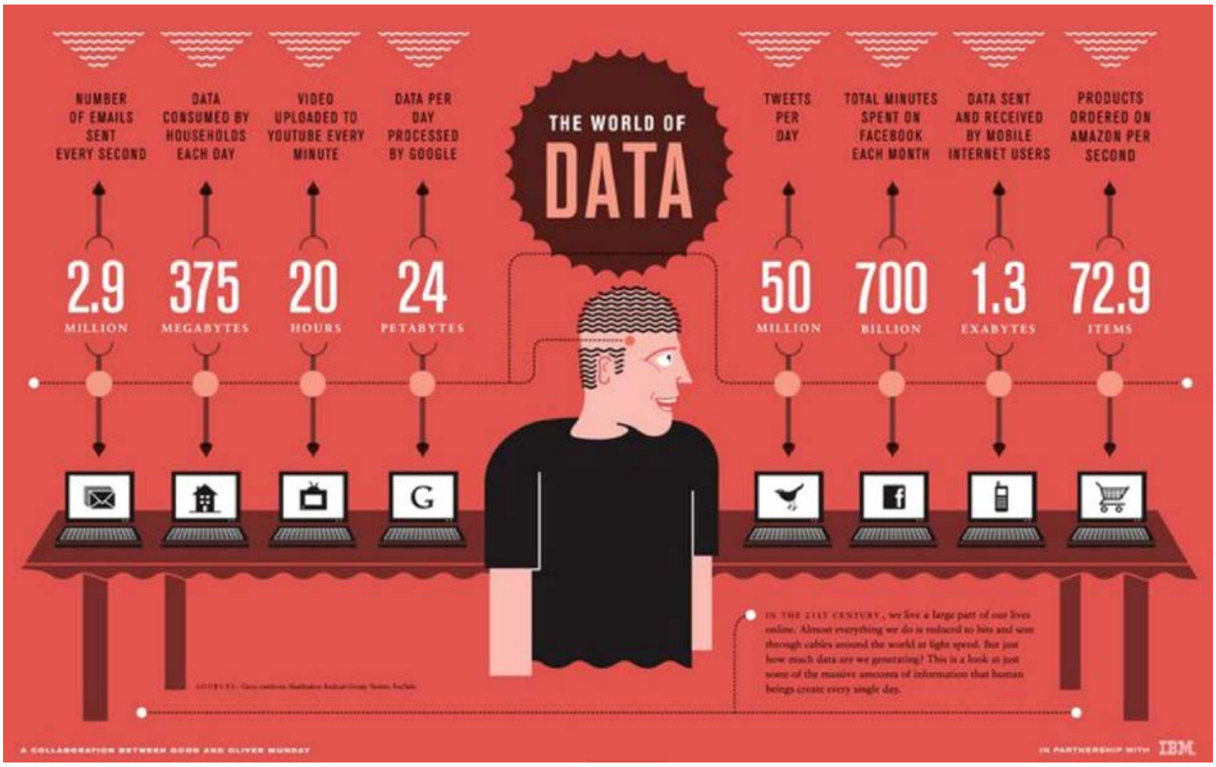


Figure 15 - The world of data

Mesh networking will soon be the major network topology. In this topology each node relays data for the network and all mesh nodes cooperate in the distribution of data within the network. Mesh networks can change shape depending on the needs. They are incredibly fast.

In recent years a new IT buzzword "Big Data" has been created, promoted and extensively used. It refers to issues related to gathering, processing and storing large amounts of (mostly) unstructured data.

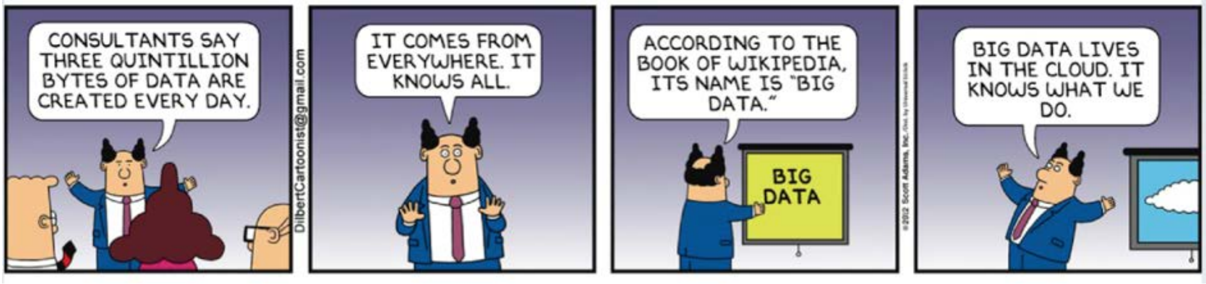

CONSULTANTS SAY THREE QUINTILLION BYTES OF DATA ARE CREATED EVERY DAY.
DilbertCartoonist@gmail.com
IT COMES FROM EVERYWHERE. IT KNOWS ALL.
ACCORDING TO THE BOOK OF WIKIPEDIA, ITS NAME IS "BIG DATA."
BIG DATA
©2012 Scott Adams, Inc.
BIG DATA LIVES IN THE CLOUD. IT KNOWS WHAT WE DO.

Figure 16 - The big data (Source: http://dilbert.com/strip/2012-07-29)

While there is a common understanding that Big Data refers to data sets that are so large or complex that traditional technologies cannot cope with storing, processing and analyzing, it is important to note that in today's use cases and applications the IT systems are pressurized even more by the fact that the datasets are constantly fed with new, unstructured data that should, in fact, be processed in real-time in order to bring the real business cases or other kinds of benefits, e.g. improved efficiency of telecommunication or power supply network, weather forecasting, stock market analytics, traffic control and optimization, etc.

For instance, sensors installed in the Internet of Things systems including devices, mobile phones, medical equipment are generating billions of data objects per second.

In 2000, the global data volume was estimated at around 800 Exabyte. Now, the world creates 5 Exabyte of new data every two days.

While relational databases and associated analytic tools are designed to interact and perform well with structured information of limited size, most of the useful information created by today's Big Data market is unstructured or semi-structured (up to 80%). This applies to photos, video, audio, XML JSON files, etc.

The discipline of science that concerns analysis of the massive data sets is also called "Data Analytics". The main goal of that analysis is to find new data dependencies and derive new information from the distributed and unstructured data sources. Based on these conclusions, we can better predict behavior relations between processes.

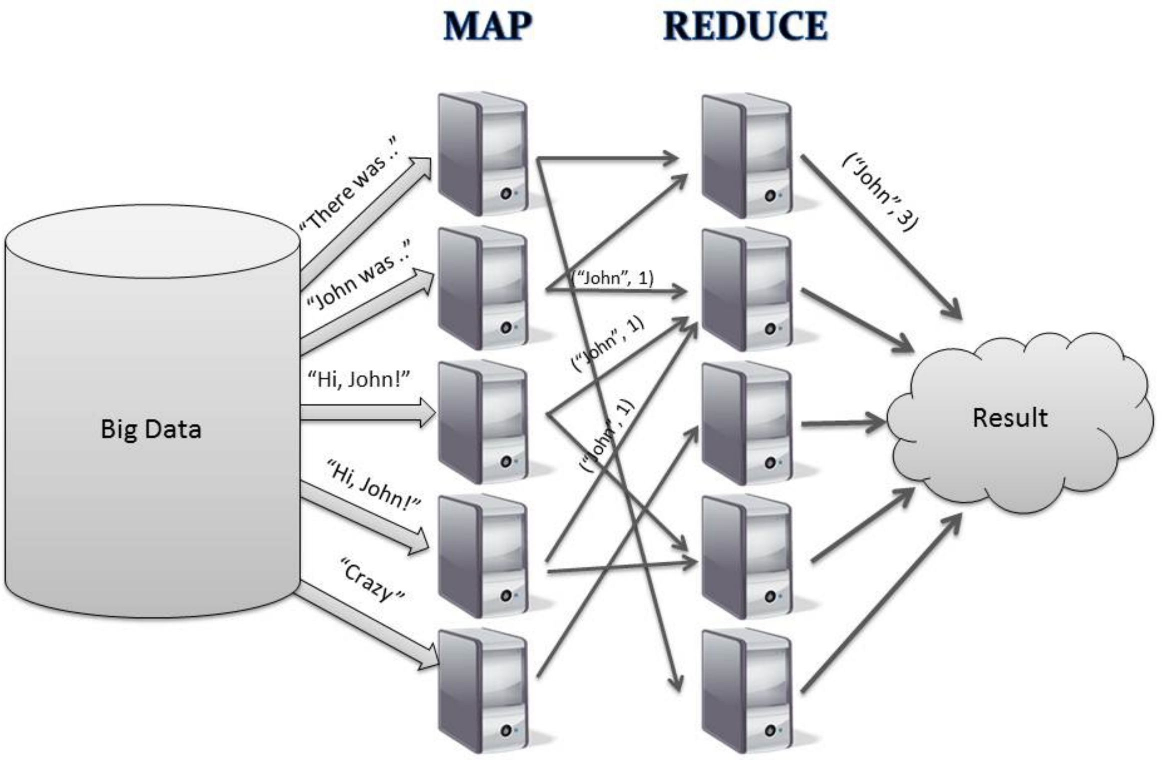
We can specify some industrial groups where Data Analytics processes are very important: financial services, telecommunications, retail, energy, healthcare and many science disciplines including earth science, biology, medicine, physics and computer science.

5.3.2 MapReduce model

Data Analytics processes are associated with some technical difficulty to overcome, such as:

- fast information growth,
- processing power consumption,
- physical storage,
- many data type issues.

The old techniques for working with information do not resolve these problems. They are too costly, time consuming and complicated. To change that situation a new methodology called "MapReduce" was designed. MapReduce was invented by Google in 2004. This is a programming model for processing and generating large data sets with a parallel, distributed algorithm on a cluster. In MapReduce, task-based programing logic is placed as close to the data as possible. There are two main components: Map and Reduce. In the Map phase, records from a large data source are divided up and processed across as many servers as possible (parallel) to produce intermediate values. After all the Map processing is done, the intermediate results are collected together and joined (Reduced) into final values. To implement the MapReduce model as a tool the Apache Pig programing language and Apache Hive, a SQL database interface, were developed.

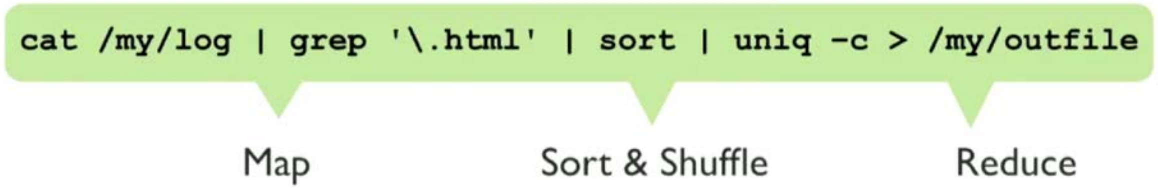


source: http://dme.rwth-aachen.de/de/research/projects/mapreduce

Figure 17 - The map-reduce model

MapReduce was a good tool for data analysis, but due to its structure it was not flexible enough as a commercial product.

MapReduce data flow can also be depicted with the following chain of shell commands:



```
cat /my/log | grep '\.html' | sort | uniq -c > /my/outfile
```

Figure 18 - Map-reduce example

5.3.3 Hadoop

For data sets analytics, we can use several IT solutions. The first of them is the Hadoop platform. It can be considered as the commercial implementation the of MapReduce model. It is a well-adopted, standard-based, open-source software framework. Initially this tool was also developed by Google. After it started becoming popular, the nonprofit Apache Software Foundation took over maintenance of Hadoop together with Yahoo. Hadoop environments are built of three basic layers:

- Application layer (end-user access layer): it provides a programming framework for software developers to apply distributed computing techniques to extract meaning from large data sets.

- MapReduce workload (management layer): also called as JobTracker. That component supplies an open-source runtime job execution engine. This engine coordinates all aspects of the environment (scheduling and launching jobs, balancing workload, failures handling).
- Distributed parallel file systems (data layer): This layer is responsible for storage and access to the information. It uses a specialized distributed file system - HDFS.

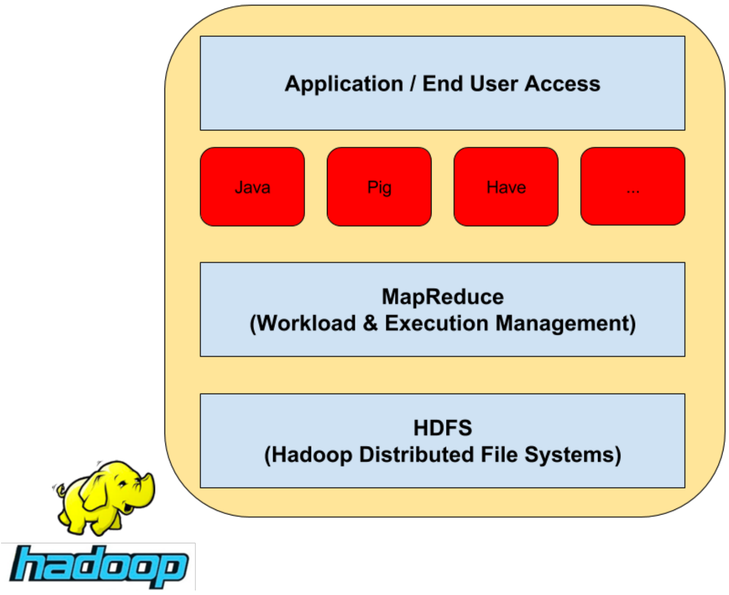

Application / End User Access
Java
Pig
Have
...
MapReduce
(Workload & Execution Management)
HDFS
(Hadoop Distributed File Systems)
hadoop

Figure 19 - Hadoop High-Level Architecture

That representation of logical hierarchy is based on full layer separation. Every layer has its own tasks to perform. Together, these layers deliver a full MapReduce implementation.

While being the most popular Big Data software stack, Hadoop also has certain disadvantages:

- Low performance - Hadoop generates lots of disk operations, which results in low system performance.
- Data isolation - when a user accesses data, he isolates them creating a data island. That island consumes hardware resources, and makes them inaccessible for the rest of users.
- Supporting multiple tenants - Hadoop has a very difficult and not flexible privilege management system, which makes providing multi-tenant configurations difficult.
- Storage restrictions - Hadoop can manage a relatively small number of information repositories (HDFS is the most popular of them).

5.3.4 Spark

Spark is getting more and more attention and maturity as the Big Data processing solution. The designers of Spark created it with the expectation that it would work with petabytes of data that were distributed across a cluster of thousands of servers (physical and virtual). Spark is a memory-optimized data processing engine that can execute operations with better performance than Hadoop. Spark reduced the amount of disk reads and writes. It can also process data on other cluster environments such as the Apache Cassandra database. Spark also supports many programming languages such as Java, Scala, Python and R.

Spark is based on the fact that in-memory processing it is always faster than interacting with a hard disk. It made Spark up to 100 times faster than earlier disk-based solutions. To implement that functionality, Resilient Distributed Datasets (RDD) were used. RDDs are in-memory structures that hold information from clients and are then processed within Spark. RDDs keep track of their history (edits and deletions) so they can reconstruct themselves if there are any failures. They are also dispersed across multiple nodes in the cluster to increase safety and efficiency via parallel processing. Data can be loaded into RDDs from about any data source a client may want:

- Hadoop File System
- Apache Cassandra DB
- Amazon S3
- Apache HBase
- Relational database

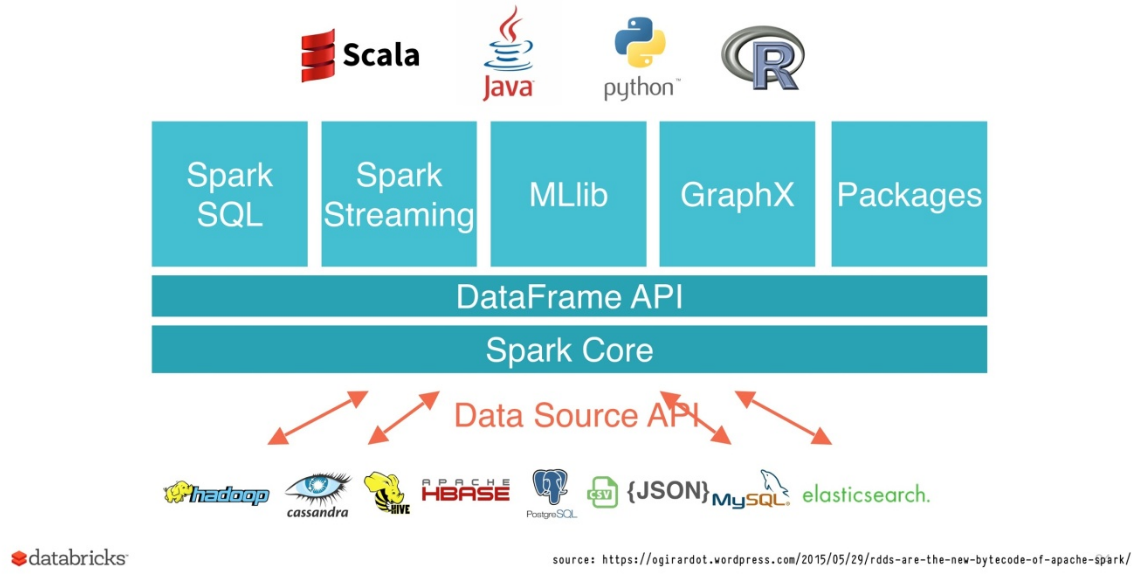
After data has been placed into an RDD, Spark permits two primary types of operations:

- Transformations - this process includes filtering, mapping or otherwise manipulating data. It also creates a new RDD - the original RDD remains unchanged. Transformations are not executed until the moment they are needed such as by an action (lazy evaluation).
- Actions - behaviors that interact with the data but do not change it, for example: counting, retrieving a specific element, and aggregating.

Spark is comprised of a collection of highly specialized components. These all work together in concert to help data scientists and developers quickly create impressively powerful applications.

The main components are:

- Core - it is the of the Spark platform. It is responsible for creating and managing RDDs as well as keeping things running smoothly across the cluster such as scheduling tasks and coordinating activities.
- Libraries - they provide additional functionalities for the Core element, for example: streaming, machine learning, SparkSQL, and GraphX
- Storage systems - Spark supports many commercial and open source file systems.



source: https://ogirardot.wordpress.com/2015/05/29/rdds-are-the-new-bytecode-of-apache-spark/

Figure 20 - The spark architecture

5.3.5 Data Analytics in HPC

Infrastructure for Data Analytics is a combination of hardware components that produce an architecture that meets specific requirements. The finest software in the world will not run without optimal processing power and storage capacity. So in developing a Data Analytics solution you start with the base, which is the infrastructure. Infrastructure must be an early part of Data Analytics system planning. Core infrastructure capabilities, critical for optimal Data Analytics infrastructure, are:

- Scalability of infrastructure — Data Analytics needs are still changing. To accommodate this growth, the infrastructure must feature scalability and resilience. Scalability provides the system ability to process a growing amount of work. Well-invented scalability makes it relatively easy to add processor and storage power.
- Parallel processing — There is parallelism in processor design and also in system level software. One obviously needs intelligently designed traffic that understands the workloads so data and processing instructions can be intelligently threaded through the hardware layer — from processor to memory to storage. A greater core can help boost performance and process more requests. The same can be said for file system parallelism. Data Analytics tools may need to be accessed in multiple ways from multiple places and at multiple rates of speed. The infrastructure design challenge is in determining where and how to store data for the best performance as data scales. High-performance parallel file systems spread the data around intelligently so it can be quickly accessed and updated at high speed.
- Low latency resources — high speed (often referred to as low latency) is becoming a real differentiator of Data Analytics systems. One of the common features of modern systems is the tremendous increase in speed, which embraces CPU speeds, RAM speeds, disk speeds, and transmission lines (optical fiber cable). For example, flash memory is very fast, and the price is coming down. It provides real benefits in accelerating Data Analytics processes. Speed can also be delivered when you optimize the hardware and software together.

Data optimization — In the world of Data Analytics, data management takes on a whole new meaning. We should decide what data we keep, whether we need to store it in a different location, what the costs are of dealing with the data, how quickly and when we need it.

5.4 Data management

Timely planning how data will be captured, collected, used, managed, stored, sustainably archived, and disseminated are all very important tasks of every researcher from the very beginning of his work. “Research data” is defined as data in the form of facts, observations, images, computer program results, recordings, measurements or experiences on which an argument, theory, test or hypothesis, or another research output is based. Data may be expressed in different formats such as numerical, descriptive, visual etc. Data can be raw or summarised / cleaned after processing. Having all this in minds data management can be defined as the processes and activities required to manage data throughout the research life-cycle for current and future research purposes and users.

A Data Management Plan (DMP) is a very useful tool for defining from the beginning of each project, what data and associated metadata and tools will be used, delivered and possible shared. Data management plans are highly valuable especially in cases where multiple users from different institutions distributed around the world are involved.

A DMP describes the data management life cycle for all datasets to be collected, processed or generated by a research project. It must cover:

- The handling of research data during & after the project.
- What data will be collected, processed or generated.
- What methodology & standards will be applied.
- Whether data will be shared /made open access & how.
- How data will be curated & preserved.

The DMP also defines what types of data services need to exist in order to fulfil their needs for the full data management lifecycle.

- Data creation: Data can be created either from some physical or natural experiment and the use of instruments or by simulation using computers.
- Data processing: Data need to be brought close to computers or supercomputers in order to be pre-processed or processed for further storage and/or use.
- Data analysis: Further analysis of processed data might be needed in order to bring the data into a desirable status or format.
- Data preservation: Data might need to be stored using services that can guarantee their long term preservation.

6 Infrastructure and Management tools

Since the publication of last year's Market and Technology Watch document [2] at least two other PRACE publications have covered the topics addressed in this chapter in considerable depth as well as breadth:

1. The HPC Infrastructure workshop #7, held at LRZ, in April 2016, the proceedings of which have been documented in [47], focused in particular on monitoring data centre infrastructure, and on analyses of power usage efficiency, both, at the level of applications and HPC machines, as well as at the level of the data centre infrastructure at large.
2. In 2016 PSNC conducted an extensive survey on data centre infrastructure monitoring (DCIM) which involved most of Europe's leading HPC centres. The results of the survey, and analyses and recommendations based on its finding, have been published in a dedicated white paper [48].

Unlike the comparable chapter 5 of D5.1 – the "prequel" of the current document - this chapter confines itself to the clarifying trends manifest in HPC centres and general conclusions with respect to the functionality of products that the market - including open source and free of charge solutions – has to offer in this context. We refer to the aforementioned documents, most notably the DCIM whitepaper, for more detailed accounts.

In HPC centres in Europe and elsewhere the improvement of energy efficiency is high on the agenda and in several ways, has an impact on the demand for management and, in particular, monitoring tools. The now manifest - slight - delays in the roadmaps for exascale compute facilities in the US are largely due to the serious commitment of the HPC community to keep the total power requirements of exascale facilities well within a maximum envelope of 30 MW, not to the inability to provide such systems per se. The required improvement of energy efficiency is expected to be achieved in large part by new compute and interconnect hardware to be designed and produced by HPC component builders and system integrators, but it cannot be achieved by improvements from the hardware supply side alone. Users, application developers and system administrators will have to do their fair share of tuning for energy efficiency improvement, and the same holds for the datacentre infrastructure.

6.1 System and application level monitoring

It is not possible to improve or tune aspects of machines and processes that one cannot adequately measure. The focus on energy efficiency, on all levels involved, first and foremost has produced a demand for monitoring capabilities at a level of component detail, and with a fine-grained time resolution, that goes far beyond what used to be needed when the purpose of monitoring was chiefly to detect malfunctioning or misconfiguration of components at an early enough stage. In addition, the variety of components that need to be monitored increases.

Chip builders and system integrators generally have shown themselves responsive to that demand, in that they continue to expand the possibilities for data extraction at the component level. The generation of HPC systems that is now put on the market by system integrators is generally equipped with sensors that make the systems aware of their compute load and that implement, or provide feasible hooks to implement, regulatory loops that minimize the energy consumption of parts that are idle or endure a lesser load. For example, clock frequencies of processors are adjustable, as are rotation speeds of blowers and/or the flow rates of the fluids that circles trough the closed circuits of liquid-cooled cabinets. Data quantifying the difference in energy consumption resulting from such adaptation can in principle be extracted at the level of allocation units to individual jobs such as CPUs or nodes and at a time scale that is appropriate for batch system and application tuning.

Developers, users and system administrators can relate such to data on application performance and/or system throughput and make policy decisions on how to run applications, or the system at large. Well, they must do so, if they want to improve energy efficiency – the hardware is not going to decide by itself what is best. Applications and systems must be designed and tuned by implementing strategies that optimize the *energy* to solution, rather than the absolute time to solution. At the HPC Infrastructure Workshop #7, Intel stated a power usage gap of about a factor three has still to be bridged before 2022 to keep an exascale system within the 30 MW envelope. With the presentation of is Global Energy Optimization (Geo) framework Intel called “all hands-on deck”. GEO aims to facilitate application runtime environments that are both energy aware and able to coordinate global energy management decisions dynamically and online, with the applications in the driver’s seat.

To facilitate analyses for energy efficiency for a given HPC facility and applications, fairly large volumes of monitoring data from heterogeneous sensors must be preserved and combined with data in the realm of system accounting and resources. The handling of the data must remain scalable. The data should for instance be easily aggregable and organisable in various hierarchical ways.

6.2 Infrastructure monitoring

It still makes sense to distinguish between HPC system monitoring and the monitoring of the supporting infrastructure: the building, the electrical and mechanical support systems. Tools that monitor the security of the premises, the quality of air and water, etc. need no tight integration with system monitoring. However, a consequence of the adoption of all-encompassing energy efficiency strategies by sites that host significant HPC facilities is, that the line between monitoring and managing the HPC facility on one side, and the monitoring of electrical and mechanical data centre infrastructure on the other side, is becoming increasingly blurred. There is an explicable demand for tools that facilitate integration of system and infrastructure monitoring, as strategies adopted to improve the system's energy efficiency may have a severe impact on the infrastructure. The compute side may be trying to drastically improve its energy footprint in ways that were never envisaged by the infrastructure side. For example, the HPC Infrastructure Workshop of 2015 [49] already documented a case that demonstrated how the ability of the HPC machine to optimize its power usage to the demand of a well-concerted HPC application, utilizing the bulk of the machine's resources, can lead to very large power fluctuations in very short time intervals – in the order of several MW in a few seconds - with which the installed electrical subsystem, and the local energy provider, failed to cope. The mere ability to analyse the causal relationships in such a case, presupposes sophisticated monitoring tools that integrate detailed data on a fine-grained time scale from the batch system side with data of comparable quality from the datacentre infrastructure side.

If anything, the necessity of the integration of data from the infrastructure side and the HPC system side is expected to increase further, as solutions to optimize the energy efficiency of the datacentre infrastructure, presented at the last HPC infrastructure workshop, now target the re-use of waste heat from HPC facilities in the core datacentre infrastructure that accommodates other hosted facilities, rather than to export it to offices or offsite heating systems. The proposed use of absorption cooling in fact boils down (pun intended) to making an HPC facility, as supplier of hot water used to dry the adsorbent in the cooling unit, into an integral part of the cooling plant for other facilities and components that need water at about 15°C, such as computer room air conditioning units (CRACs) and rear door heat exchanges on air-cooled cabinets.

Even without such novel dependencies between systems and infrastructure in place, the desire to attribute the proper portion of energy usage and other infrastructural costs to identifiable machine units (nodes, racks) that are allocated to a specific project or job for a specific time interval, requires integration of infrastructure monitoring and batch system monitoring data. The motivation to gain insight in the "infrastructural cost per job" can vary. It may stem from the datacentre's business model - the proper portion of overhead has to be billed to clients. It may also merely serve to heighten user awareness, while accounting is based on the wall clock time of resources usage, irrespective of the intensity of power usage during that time interval.

6.3 Integration of system and infrastructure monitoring in HPC centres

Various accounts by sites presented at the last HPC infrastructure workshop as well as the results of the DCIM survey clearly show that HPC centres are invariably actively pursuing such integration of monitoring. Well-structured "dashboards", providing a bird's eye view of integrated data from many subsystems, are highly sought after. The DCIM survey also shows the supply side of the market currently offers centres a very rich portfolio to choose from – but only of partial solutions. To cover all monitoring requirements of the datacentre infrastructure side, sites typically use three or more independent platforms. Customization and integration between such platforms – as far as is possible – is done mainly by the sites themselves and is costly in terms of manpower needed to develop and maintain the systems.

Part of this is due to the enormous variation of equipment – also in age - deployed and the way it is connected to the network. The DCIM report also shows that sites usually implement a dedicated closed network on which all monitoring and control of datacentre infrastructure equipment takes place. Remote access, if possible at all, is usually limited and only granted though one or two well-controlled “stepping stones”. While this setup may not be conducive to easy integration of data, in many cases it simply necessary to protect system integrity. Datacentres, and many of their electrical and mechanical subsystem components, have a much longer lifespan than the IT equipment they host and support. Interfaces that give access to equipment control functions may be not well shielded and be practically inseparable from the interface needed for mere monitoring of performance and energy usage. The ability to query equipment state may have been envisaged to be used only by staff and hardware field engineers that at the same time need an interface to repair or reconfigure a malfunctioning device. Access to the monitoring functions – or to the data that those provide – may now be deemed relevant for a wider audience, but that audience should not gain access to control functions.

For some infrastructural equipment, continuous monitoring of equipment state at high frequency may not have been envisaged at all. Some centres improvise and place additional sensors and meters that enable them to get a fair estimate of such devices after all. This tends to lead to quite eclectic, very site specific, data acquisition procedures.

7 System tools and programming environment

For dynamic resources (e.g.VM), the Big Data community provides new system tools that are targeted to web developers and administrators. The programming environments are evolving from compiled languages to interpreted ones with powerful libraries.

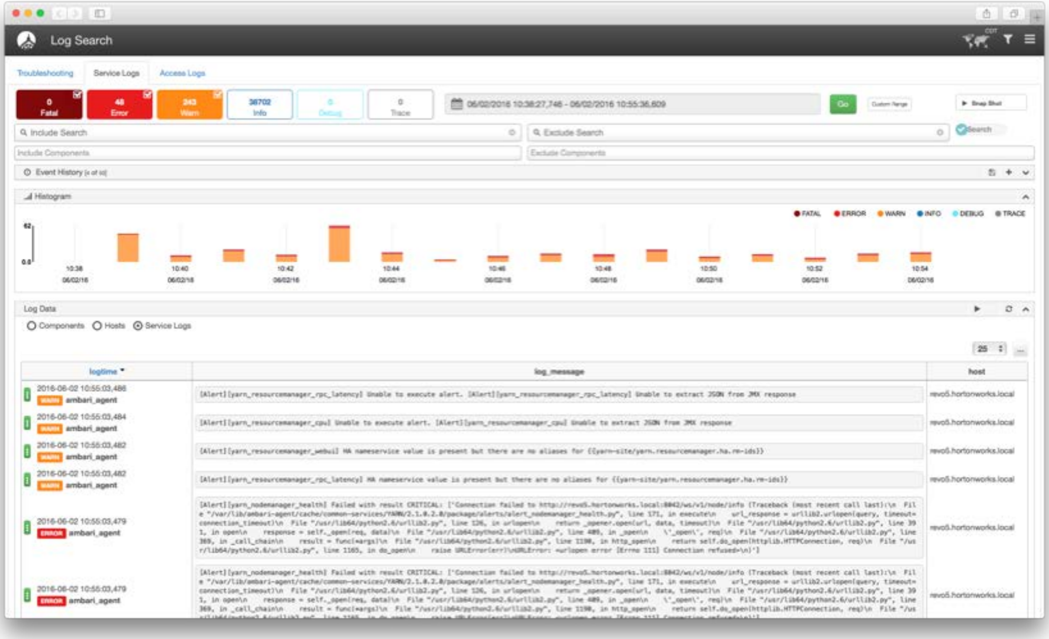
7.1 System / usage oriented management software

Intuitive tools are already in production for managing millions of virtualized resources and should be considered for managing HPC clusters.

7.1.1 Apache Ambari

Apache Ambari [50] allows provisioning, managing, monitoring and securing resources.

It comes with an easy installation method then it provides a wizard-driven and automated cluster provisioning. The cluster upgrades can be done in an automated rolling or in express mode.



logtime	log_message	host
2016-06-02 10:55:03,486 WARN ambari_agent	[Alert][yarn_resourcemanager_rpc_latency] Unable to execute alert. [Alert][yarn_resourcemanager_rpc_latency] Unable to extract JSON from JMX response	revo5.hortonworks.local
2016-06-02 10:55:03,484 WARN ambari_agent	[Alert][yarn_resourcemanager_cpu] Unable to execute alert. [Alert][yarn_resourcemanager_cpu] Unable to extract JSON from JMX response	revo5.hortonworks.local
2016-06-02 10:55:03,482 WARN ambari_agent	[Alert][yarn_resourcemanager_webui] HA nameservice value is present but there are no aliases for {{yarn-site/yarn.resourcemanager.ha.rm-ids}}	revo5.hortonworks.local
2016-06-02 10:55:03,482 WARN ambari_agent	[Alert][yarn_resourcemanager_rpc_latency] HA nameservice value is present but there are no aliases for {{yarn-site/yarn.resourcemanager.ha.rm-ids}}	revo5.hortonworks.local
2016-06-02 10:55:03,479 ERROR ambari_agent	[Alert][yarn_nodemanager_health] Failed with result CRITICAL: ['Connection failed to http://revo5.hortonworks.local:8042/ws/v1/node/info (Traceback (most recent call last):\n File "/var/lib/ambari-agent/cache/common-services/YARN/2.1.0.2.0/package/alerts/alert_nodemanager_health.py", line 171, in execute\n url_response = urllib2.urlopen(query, timeout=connection_timeout)\n File "/usr/lib64/python2.6/urllib2.py", line 126, in urlopen\n return _opener.open(url, data, timeout)\n File "/usr/lib64/python2.6/urllib2.py", line 391, in open\n response = self._open(req, data)\n File "/usr/lib64/python2.6/urllib2.py", line 409, in _open\n \'_open\', req)\n File "/usr/lib64/python2.6/urllib2.py", line 369, in _call_chain\n result = func(*args)\n File "/usr/lib64/python2.6/urllib2.py", line 1190, in http_open\n return self.do_open(httplib.HTTPConnection, req)\n File "/usr/lib64/python2.6/urllib2.py", line 1165, in do_open\n raise URLError(err)\nURLError: <urlopen error [Errno 111] Connection refused>\n)']	revo5.hortonworks.local
2016-06-02 10:55:03,479 ERROR ambari_agent	[Alert][yarn_nodemanager_health] Failed with result CRITICAL: ['Connection failed to http://revo5.hortonworks.local:8042/ws/v1/node/info (Traceback (most recent call last):\n File "/var/lib/ambari-agent/cache/common-services/YARN/2.1.0.2.0/package/alerts/alert_nodemanager_health.py", line 171, in execute\n url_response = urllib2.urlopen(query, timeout=connection_timeout)\n File "/usr/lib64/python2.6/urllib2.py", line 126, in urlopen\n return _opener.open(url, data, timeout)\n File "/usr/lib64/python2.6/urllib2.py", line 391, in open\n response = self._open(req, data)\n File "/usr/lib64/python2.6/urllib2.py", line 409, in _open\n \'_open\', req)\n File "/usr/lib64/python2.6/urllib2.py", line 369, in _call_chain\n result = func(*args)\n File "/usr/lib64/python2.6/urllib2.py", line 1190, in http_open\n return self.do_open(httplib.HTTPConnection, req)\n File "/us	revo5.hortonworks.local

Figure 21 - The Ambari dashboard

Thanks to a centralized security setup, it reduces the complexity to administer security across the platform and simplify the hardening of this privileged part of the cluster. As an example of ease of use, Ambari has an automatic setup for Kerberos. Ambari Views offers the possibility to customize the user interface by assigning different roles for each kind of administrator.

As shown in the dashboard above, there is a full visibility into cluster health. Moreover, some alerts are already predefined based on operational best practices. The advanced metrics visualization is made with Grafana. It can seamlessly fit into an existing environment, as the administrators are able to bring custom services under management via Ambari Stacks.

To determine what exactly happened, Ambari log search with Solr (Elastic search like) can be executed on a cluster, possibly dedicated to this task, working with the Log Warehouse. It will give insights to the HPC administrators as shown in the Log search screenshot.

7.1.2 Singularity (container suitable for HPC)

Thanks to Singularity [51], scientists can control their whole software stack and have a better reproducibility of their simulations on different HPC clusters. Its overhead is fair as it is a starting project, we can see it on the genomic program performances below.

It has single file based container images which facilitate distribution, archiving, and sharing with other research teams. There is no system, architectural nor workflow changes necessary to integrate on HPC s.

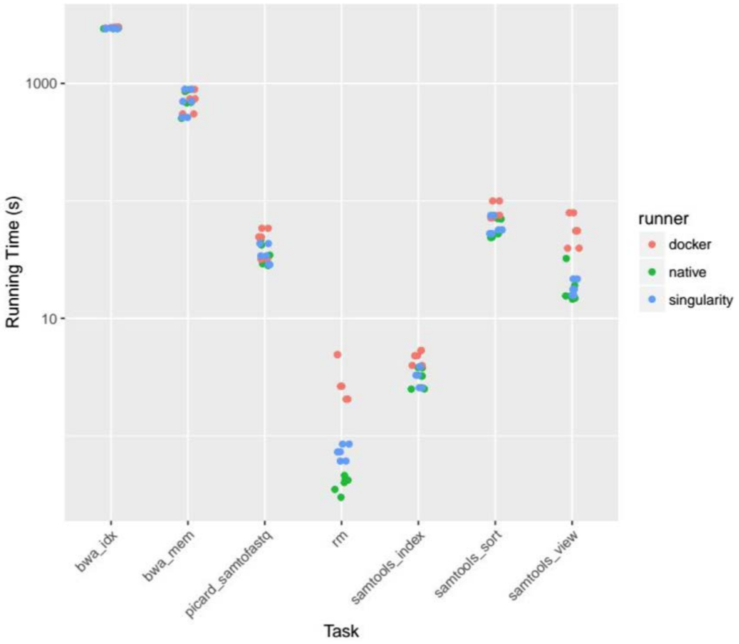


Figure 22 - enomic programs performances in Singularity container

User's privileges within the container are the same as the outside of it. For better security, there is no root owned container daemon. Finally, its integration with resource managers, InfiniBand, GPUs, MPI, file systems is simple and it supports the widely-used architectures that are x86_64, PPC and ARM.

7.1.3 Intel GEOPM

Global Extensible Open Power Management [53] is a framework developed by mostly Intel staff for optimizing the power consumption of a compute job when run on a distributed system. MPI applications can use a library to participate in the control of the power policy. Unmodified applications can have a standalone control application running alongside, this then queries the hardware registers in the CPU. GEOPM takes advantage of the fact that modern CPUs have many cores available and runs software on all compute nodes used for a job.

While the application is running, the performance of each rank is monitored so that power is used on the compute nodes that need it the most. Nodes that are falling behind can get a power boost while still maintaining a global power envelope, for example.

GEOPM has previously been covered by WP5 in D5.4, section 6.1, page 14.

7.2 Programming environment

This section contains tools both for developing applications, but also for run-time support.

7.2.1 Charm++

Object oriented programming model for C++ uses parallel objects that run concurrently and communicate using normal C++ methods, and while this ties the framework to C++ applications it also makes it more natural to use for C++ developers. The objects can migrate between compute nodes along with data if a serialization method is provided in the class. It allows the programmer to combine objects into collections and invoke C++ methods on an entire collection at once. A runtime environment handles the movement and placement of objects within a job. One major popular application using Charm++ is the molecular dynamics code NAMD.

Charm++ can be combined with MPI and OpenMP. For the developer a parallel debugger, Charm Debug, and tracing/profiling tool are available.

Charm++ is being commercialized by Charmworks[2], founded by researchers from the University of Illinois at Urbana-Champaign who have been developing Charm++.

7.2.2 ARM development software

The development tools available for x86-based systems are often seen as one of their advantages, with for example Intel providing a wide range of compilers and performance tools for their CPUs. For ARM systems to gain a foothold in the HPC market an equivalent level of tool support is beneficial, which ARM seems to have recognized. In recent years it has both been developing optimized math libraries for HPC and has acquired the development tool vendor Allinea. All software from ARM targeted for the HPC market is intended to be used on systems with 64-bit ARMv8 processors.

ARM Performance Libraries is the math library offering. Much like Intel MKL this provides optimized BLAS, LAPACK and FFT routines.

ARM Compiler for HPC is a bundle that contains both a C/C++ compiler and the ARM Performance Libraries. The compiler is based on LLVM and targets ARMv8-A.

ARM SVE Compiler for HPC is a version of the compiler suite with support for the new vector extensions.

ARM Code Advisor will be a tool for both static and dynamic analysis/profiling, but is not released yet and only available to registered beta testers. CoreSight traces can be enabled in the ARM debugger, and the **Streamline** tool is available from the SoC development toolset for parallelization and power usage monitoring.

Notably absent is Fortran support in the compiler suite, which is due to LLVM not having a Fortran frontend yet. Currently “Flang”, an adaptation of the PGI Fortran frontend, is being developed for LLVM. Once this is complete the ARM compiler will probably include Fortran support. NAG and PathScale provide commercially supported Fortran compilers for ARMv8-A. GNU GCC is the open source option with its Fortran frontend.

7.2.3 High-productivity programming languages

In recent years, alternative programming languages have become more popular for HPC applications, complementing the old C/C++/Fortran mainstays while still being able to leverage highly optimized libraries written in those languages. A common term for these is “high-productivity programming languages”, with Python, Julia and R being examples of such languages. One thing these have in common is that they can be used both in batch and interactive modes.

Python For many years the Python language has been popular as a scripting language, but it has steadily increased its computational capabilities with packages such as Numpy and Scipy. Getting these packages built against an optimized math library has not been without its issues.

[2] http://hpccharm.com/

In recent years, the Anaconda distribution has become a popular way to install a predefined Python stack. It is a binary distribution of Python, with prebuilt packages linked against math libraries like Intel MKL. The Anaconda framework itself also supports other languages, but was originally created for Python. One current issue with Anaconda is that it is geared towards single-user installations, with shared cluster-wide installations needing some manual fixes.

R is a language suited for statistical purposes, and it fits well with the trend of doing data analysis on big data sets. Based on the S language from Bell Labs, R was originally an open source implementation of S, but has now eclipsed its inspiration. It is mainly used for data analysis. Starting with version 2.14 in late 2011 the language has provided support for parallel programming, mainly for multicore but also with some basic support for MPI. R also has packages to provide an interface to Hadoop.

Julia is a fairly new programming language that aims to combine features from Python, R and other languages. It also has some traction as a replacement for Matlab. Interactive usage is supported primarily through the use on IJulia, which is based on the notebook paradigm popular in mathematical software. These notebooks can have interactive widgets for data visualization. IJulia uses the Jupyter notebook framework, and creates a web application that needs a web browser to access, so this doesn’t fit directly into a cluster where compute nodes are not accessible from the outside.

8 EU Projects for Exascale and Big Data including PCPs and related prototypes

PRACE-4IP Deliverable D5.1 “Market and Technology Watch Report Year 1” contains a thorough description of EU projects (Horizon 2020, but also FP7, as well as Eureka ITEA2 and ITEA3) in the area of HPC, including exascale objectives. It also describes organization of the R&D FETHPC and CoE programmes oriented and monitored in the framework of the HPC cPPP (Chapter 2).

This section is focusing on updates and evolutions of these efforts.

8.1 7th Framework Programme

Previous funding on Exascale efforts (predating the cPPP setup) amounted to more than 50M€ for 8 projects [54].

In addition to software development, 2 tracks of hardware prototypes arose from resp. DEEP/DEEP-ER [55] and MontBlanc [56] 1&2 projects:

- The final DEEP prototype system consists of a 128-node Eurotech Aurora Cluster (Intel Xeon+Infiniband) and two distinct prototypes for the Booster:
    - A 384-node system built by Eurotech from custom-engineered dual-node cards in the Aurora blade form factor – the DEEP Booster with aggregated performance around 500 TFlop/s
    - A smaller 32-node prototype built by University of Heidelberg and Megware based on the latest ASIC implementation of EXTOLL
- The current MontBlanc prototype consists of two racks of blade servers that make use of the Bull bullx B505 blade server, and BSC chose the Exynos 5 ARM chip from Samsung, which has two Cortex-A15 cores running at 1.7 GHz, for a CPU on each node.

This MontBlanc 2 machine has eight BullX blade enclosures, for 72 carrier blades with a total of 1,080 CPU-GPU cards in two racks, yielding 34.7 teraflops in a 24-kilowatt power budget.

8.2 H2020 FETHPC projects

All 19 R&D projects stemming from the FETHPC Call in 2014 are up and running (plus two Coordination and Support Actions, EXDCI and Eurolab4HPC, for ecosystem development and extra prospective) [57].

Some documentation on the progress and potential of these projects can be found at [58]. ETP4HPC and EXDCI have produced Handbooks at the occasion of BoF sessions at Super Computing (2015 and again in 2016), with up-to-date detailing of the European HPC Technology Projects within the European HPC Eco-system [62].

8.3 Centres of Excellence for Computing Applications

A ninth Centre of Excellence has joined the initial group of eight selected in 2015 [59]:

- EoCoE - Energy oriented Centre of Excellence for computer applications;
- BioExcel - Centre of Excellence for Biomolecular Research;
- NoMaD - The Novel Materials Discovery Laboratory;
- MaX - Materials design at the eXascale;
- ESiWACE - Excellence in SImulation of Weather and Climate in Europe;
- E-CAM - An e-infrastructure for software, training and consultancy in simulation and modelling;
- POP - Performance Optimisation and Productivity;
- COEGSS - Centre of Excellence for Global Systems Science.
- CompBioMed [60]- started in 2016 - Computational Biomedicine

8.4 Next Work Programme 2018-2020

ETP4HPC, Centres of Excellence (in the framework of the HPC cPPP) and other stakeholders were mobilized in 2016 to elaborate recommendations for the next part of Horizon 2020 HPC efforts. This is work in progress, including, but not limited to, ETP4HPC SRA (Multi Annual Roadmap) [61]. The scope of reflections encompasses the follow-up of technological (hardware and software) R&D, a concept of “integrated demonstrators” building on H2020-funded R&D, the future of Centres of Excellence, and more ecosystem development support.

8.5 Pre-Commercial Procurements and related prototypes

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

The PRACE-3IP PCP opened a call for tender in November 2013. Currently, the PCP is in phase III, for which the following suppliers had been awarded a contract: ATOS/Bull SAS (France), E4 Computer Engineering (Italy) and Maxeler Technologies (UK). During this final phase, which started in October 2016, the contractors will have to deploy a pilot system with a compute capability of around 1 PFlop/s, to demonstrate technology readiness of the proposed solution and the progress in terms of energy efficiency, using high frequency monitoring designed for this purpose. The access to these systems will be open to PRACE partners, after the PCP team have performed their evaluations.

The HBP PCP finished on 31.1.2017. To ensure that suitable solutions for realizing HBP's future High-Performance Analytics and Computing Platform will exist, the project published a tender for a PCP in April 2014 focussing on R&D services in the following areas: integration of dense memory technologies, scalable visualization as well as dynamic management of resources required for interactive access to the systems. In phase III Cray and a consortium consisting of IBM and NVIDIA were selected. These contractors implemented their proposed solutions and evaluation is ongoing since the pilot systems had been installed in summer 2016. JUELICH will continue to keep the pilot systems in operation to keep the solutions available to the HBP project. As such the systems have been integrated into HBP's High-Performance Analytics and Compute Platform (HPAC).

9 Major paradigm shifts in HPC market and technologies

9.1 Consolidation in HPC market

During the last two to three years the HPC market has seen big traditional HPC companies being acquired by other traditional HPC companies with a long history and big market share in the HPC market. In the following paragraphs, we present some information on such acquisitions, the results of whose, where also visible in the last Super Computing conference.

9.1.1 ATOS acquired BULL

In May 2014, Atos, an international information technology services company, and Bull, a European partner known among others for enterprise data services, together announced the intended public offer in cash by Atos for all the issued and outstanding shares in the capital of Bull [63].

Atos offered €4.90 per Bull’s share in cash, representing a 22% premium over the Bull’s closing price (€4.01) on Friday 23 May 2014, the last trading day before this announcement of the acquisition, and a 30% premium with respect to the 3-month volume weighted average share price (€3.77). The offer also targeted the outstanding Bull’s OCEANEs at €5.55 per OCEANEs [63].

The offer valued the fully diluted share capital of Bull Group at approximately €620 million. The two companies had complementary technologies and skills in critical segments which could offer acceleration in the areas of cloud, big data and cybersecurity solutions.

Even before the merger with SGI, some of their systems were sold by HPE. At the beginning of 2016 HPE announced an OEM agreement with SGI to sell the SGI UV system, under the name Integrity MC990 X, for the large shared memory market. HPE previously had an eight-socket system, DL980, but this was not refreshed beyond G7 in the DL range. Today HPE seems to consider scale up systems as part of the more enterprise focused Integrity range.

Despite the merger, the www.hpe.com and www.sgi.com websites still operate largely independent of each other at the time of writing, presenting a portfolio of products of HPE and SGI respectively, including products deemed to be discontinued in near future. The product naming has not been unified within a common naming scheme, the product presentation and placement has not been unified towards the public.

Support considerations regarding the merger remain a concern for SGI customers at the time of writing. As per signals emanating from high-ranking representatives of both companies, the individual legal entities of the SGI are expected to continue existence for the next few years, retaining the original support teams, procedures and structure. However, no concise and definite policies were published by the management as yet.

9.1.2 HPE acquired SGI

In November 2016 Hewlett Packard Enterprise (HPE) has acquired a high-performance computing company with a long history SGI - in a deal worth $275 million [65]. HPE announced it is buying this high-performance solutions company for computing, data analytics, and data management, in a deal worth about $275 million.

The acquisition strengthens HPE's position in the $11 billion HPC segment and in the data analytics segment. Plus, it expands the company's presence in key vertical markets for HPC, such as government, research, and life sciences.

It needs mentioning that HPE originates from the split of Hewlett-Packard, to HP Inc. and HPE, in 2015. The split created HP Inc., which owns the PC and printer business, and HPE, which owns the enterprise hardware, software, and services business.

In May 2016, HPE announced it would spin off its Enterprise Services business and merge that with another IT services outsourcing giant Computer Sciences Corp., more commonly known as CSC, creating a company worth $26 billion.

SGI eventually rebranded itself from Silicon Graphics to SGI, and also rebranded itself as a supercomputer vendor. In 2009 SGI sold itself to Rackable Systems. Rackable, which sold computer hardware and software designed for data centre deployment, recognized that SGI had a more recognizable brand, and took that name.

9.1.3 DELL acquired EMC

On September 2016, Silver Lake-backed Dell Inc. completed its $60 billion deal to acquire EMC Corp. This was considered the largest technology merger in history [66].

The new company, is named Dell Technologies, aims to be a one-stop shop for information technology sold to business. With $74 billion in revenue, Dell Technologies will be the world’s largest privately controlled technology company. As the corporate computing market shifts to newer technologies and cloud services, Dell anticipates consolidation in the market for conventional servers and storage hardware.

9.2 Artificial intelligence

Artificial Intelligence (AI) is a broad concept of machines capable of intelligent behavior, able to perform tasks that normally require human intelligence. It includes multiple subfields like control theory, search and optimization, machine learning (ML), reasoning, etc.

In the past years, ML has been the focal point of research, attracting attention due to breakthroughs in various recognition tasks – may it be voice, image or video. Other successful applications include autonomous driving, recommendation engines, internet advertising, news clustering – related stories, handwriting recognition, questionable content identification, automatic closed captioning, machine translation, and so on. Looking at how the definition of ML changed over the past fifty years, it is noticeable that the field became more data-driven as it evolved from giving "computers the ability to learn without being explicitly programmed" (A. Samuel, 1959) to constructing "computer programs that automatically improve with experience" (T. Mitchell, 1997).

The combination of a certain set of operations and a vast quantity of data fueled the research and development of specialized hardware, especially stream processing, as some algorithms, most prominently the deep learning ones, feature a high degree of parallelism. All this lead to widespread usage of graphical processing units (GPUs), with the latest NVIDIA Pascal architecture being the frontrunner at this point. It features key concepts like high bandwidth memory, half-precision floating-point and a new high-speed interconnect - NVLink. Competitor AMD features the Next-Generation Compute Unit (NCU). Although not unique to it, the architecture benefits from some very interesting features like high bandwidth memory and half-precision floating-point. On the mobile - embedded side, ARM is offering the Bifrost architecture, with quad vectorization and 16-bit floating-point support. This is present in the Mali GPUs and is already integrated in a number of handheld devices powering the inference step of applications such as image, text and voice recognition. Apart from the usage of GPUs, in the past years, new architectures emerged, such as NYU's NeuFlow embedded hardware for real-time vision, with optimizations for vision and convolutional network algorithms. After the acquisition of Nervana in 2016, Intel announced a new architecture that will be launched in 2017, specialized for ML tasks – the "Lake Crest" chip. They also announced "Knights Crest", which should be a combination between Nervana's ASIC that optimizes certain neural networks operations and the architectures present in Knights Landing and Xeon.

Abstracting from the hardware layer, there are several accelerated libraries of primitives for the most common mathematical operations and neural network routines, like NVIDIA cuDNN, Intel MKL, Apple BNNS, AMD MIOpen or the community-developed NNPACK. The main goal of these libraries is to provide high-performance implementations that can be leveraged by higher-level frameworks in multi-core and multi-node environments. An interesting initiative is AMD's ROCm open-source HPC/Hyperscale-class platform for GPU computing that is programming-language independent, and relies at runtime (ROCr) on the Heterogeneous System Architecture (HSA) Runtime API.

At a higher framework level, the competition is quite fierce. There are many options, open-source or not, featuring domain specific high-level routines optimizations with bindings for most common languages (C/C++, Python, R, Lua, Scala, Go, Javascript etc.). A common future goal for all these contenders is efficient multi-node scaling with specialized, high speed, distributed communication protocols over InfiniBand like fabrics. The most important frameworks available are Google's Tensorflow, Berkley's/Intel's Caffe, University of Montreal's Theano, Microsoft's CNTK, NVIDIA's Digits, Facebook's Torch, Nervana's Neon, Apache's Mahout, Baidu's Paddle, Numenta's Nupic, and University of Washington's/Carnegie Mellon University's mxnet.

9.3 Quantum computing

This is the first-time Quantum Computing (QC) enters in PRACE technology watch with a specific paragraph, and this is indeed a first sign this technology is growing in importance for the HPC world.

In a nutshell, Quantum Computing is the usage of a controlled quantum system to compute solutions of computational problems. The intrinsic property of the superposition of states of a quantum system is the key feature that allows a quantum computer to compute solutions of complex problems at a speed that is beyond comparison with a typical digital computer, e.g. a quantum computer can in principle solve a problem with factorial complexity with a single instruction.

Quantum Computer (QC) concept is not new, but since a few years ago Quantum Computers were only the subject of academic and speculative activities, mainly for two reasons: the performance growth of digital computers was exponential and keeping the pace of the growth rate of the demand, and the lifetime of available q-bits technology was too small for a real device.

Today the situation is different, we are assisting at the end of downscaling of silicon transistors, which are at the base of digital computers, and the lifetime of today q-bits has become much longer, making the practical realization of a QC possible.

This combination is driving today the investments of industries well established in the digital computing market, embodied by the giants Intel, Microsoft, Google and IBM, among others. All have their own proof of concept in QC technology and few are starting to offer access to them (e.g. IBM offer access to its 7 Q-bits quantum computer) for small test-drive.

Beside the above giants of the computer industry, there are also some new companies, which were born specifically to commercialize quantum computers, or to offer competences and services to help building QCs (e.g. the capability to implement a q-bit with a specific technology process).

Among those "native" QC vendors, the most successful one is D-Wave, whose technology is able to implement thousands of q-bits in a QC, even if its design relax the constraints to have a single coherent quantum state across all q-bits, but integrates many locally coherent domains.

Other start-ups, like 1Qbit [67] offer services and competences on the development of software for quantum computers, for those that will like to plan the development of new applications to be ready when QCs will hit the markets. Most of these companies (like MagiQ [68]) have also know-how in domains like cryptography that are well suited for quantum logics, so in many cases QC are only a long shot goal for their customers that are interested in knowing more about related problems (cybersecurity) on digital computers or digital communications as well.

Clearly, HPC is among those fields where quantum computing could be used in synergy with digital computers to speed-up complex numerical kernels. Traditional HPC systems could take care of all the instructions, which cannot be executed on the quantum computer (typically I/O, and control code) offloading to the quantum engine by means of a dedicated library of computational kernels that can be reformulated using the quantum logic.

Then, most probably, the quantum computers will not replace the digital computers but will integrate them. In principle one quantum computer could be shared among different applications and users (probably being accessible through a cloud), but for HPC the most probable configuration is the shipping of a quantum engine together with the HPC system. Today, commercially available quantum computers are quite expensive (in the order of 10M€) but the price is less than the price of high-end HPC systems.

The quantum computer is a disruptive technology affecting both the HW configuration of HPC systems and applications, and a quantum-enabled HPC world will look completely different from what we know today. Programming paradigms will need to incorporate quantum logic, facility management and system administration functions will need to acquire new competences to manage the quantum computers.

Finally, given the growing relevance of QC, the European Commission recently proposed to make €1 billion available for a Quantum Flagship, a large-scale European research program for quantum technology, and build new units for Quantum and HPC technologies.

9.4 Neuromorphic computing

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

In this section, we highlight some selected neuromorphic computing devices.

9.4.1 BrainScaleS

The BrainScaleS neuromorphic system has been developed at the University of Heidelberg [71]. It uses analogue circuits to implement models of neuronal processes. A special feature is its wafer-scale integration, which allows for very fast communication between the neurons within a wafer. Both features allow for simulations to run several orders of magnitude faster compared to biological speeds. A BrainScaleS wafer consists of 48 reticles, each of which holds eight High-Count Analogue Neural Network (HiCANN) dies. Each of these dies implements 512 neurons and over 100 000 synapses. There are two levels of communication: one within the wafer and one between the wafer. The latter is realised through serial links implemented in FPGAs.

Based on the BrainScaleS architecture a system has been built at University of Heidelberg, which is part of the Neuromorphic Computing Platform of the Human Brain Project [HBP-NCP]. The system has been mainly used for brain research so far.

9.4.2 SpiNNacker

SpiNNaker (Spiking Neural Network Architecture) is an architecture based on simple ARM9 cores mainly developed at Manchester University [69]. The processor is based on a system-on-a-chip design integrating 18 cores as well as the network logic including 6 communication links. The architecture is optimised for brain simulations and for this reason was able to discard usually applied design principles as memory coherence, synchronicity and determinism. Point-to-point communication happens through unreliable fire-and-forget transmissions of small packets.

The goal of the project is to realise a very large system comprising 1,036,800 total ARM9 cores as part of the Neuromorphic Computing Platform of the Human Brain Project [70]. The architecture has been used for modelling of neuronal networks for brain research as well as for robotics interaction up to now.

9.4.3 TrueNorth

TrueNorth is a reconfigurable processor developed by IBM comprising 1 million artificial neurons and 256 million artificial synapses which are organised in 4096 neurosynaptic cores [72]. Like SpiNNacker the design is digital but asynchronous (except for a clock running at an extremely low frequency of 1 kHz). The chips can be connected directly together to form larger systems. A scale-up configuration has been realised integrating 16 chips on a single board. For an alternative scale-out configuration single-chip boards are interconnected via a 1-gigabit Ethernet network. The design is heavily optimised for power efficiency. A single TrueNorth chip consumes only 70 mW.

A full ecosystem around the TrueNorth hardware is growing and meanwhile in use at more than 30 universities, government agencies, and labs. It is applied to a growing number of different areas including signal processing (e.g. video tracking, supernova detection), robotics, neural circuit modelling, and optimisation.

9.5 Heterogeneous systems

Heterogeneous systems are presented as a possible solution to increase performance or reduce overall energy consumption by having more than one kind of processor in the same system. In general, it is possible to distinguish two big families of heterogeneous architectures: architectures where heterogeneity is expressed inside the node with the computing power coming from computational units with different micro-architectures; and architectures where heterogeneity is expressed at the system level with homogenous nodes featuring different micro-architectures integrated in the same network topology.

Heterogeneity is gaining importance thanks to the availability of specialized hardware, optimized for a specific workload, and the increasing diversity of workloads as compared to traditional scientific and technical computing, requiring different system design points.

The main purpose of these systems is to achieve the maximum performance of all the mixed systems available, but it implies a new variety of issues that need to be solved, such as binary incompatibilites, different endianness, different memory layouts, different calling conventions, library availability, interconnection, etc.

As such problems arose, the Heterogeneous System Architecture (HSA) Foundation was created. It works on the Heterogeneous System Architecture, a set of royalty-free computer hardware specifications, as well as open source software development tools needed to use HSA features in application software. The HSA Foundation ambition is to develop and define features and interfaces for various types of computer processors, including CPUs, graphics processors, DSPs; as well as the memory systems that connect these. The resulting architecture, HSA, intends to make it easier to program parallel systems built from heterogeneous combinations of these devices. Current members of the foundation are listed in [73]. It worth's noticing that neither Intel nor IBM are members of the HSA at the time of writing of this document.

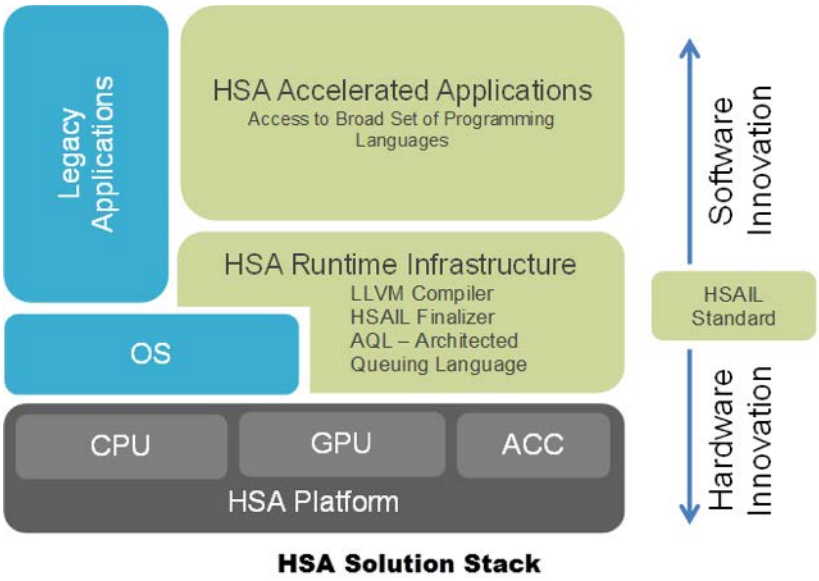


Figure 23 - HSA Solution Stack

Heterogeneous systems in the HPC ecosystem are a reality, however there is a real limitation on the number of architectures integrated in the same system. Besides this, most of the scientific applications aren't completely ported to support these complex heterogeneous systems. Efforts are made in order to achieve this, but it's a complex task which requires some time.

Concerning architectures with heterogeneous nodes, the survey of latest announcements shows a big push towards many kinds of accelerators that comes from the needs of different fields. By taking a broader view including HPDA and the big hyperscalers, we see that:

- Artificial Intelligence (IA) and Deep Learning (DL) worlds are pushing towards GPU based solutions, at least for the training part. This becomes the bread and butter of NVIDIA (the different declinations of Pascal, its Parker SOC, the forthcoming Xavier). Note that Intel is following suit with its Knights Mill (KNM), a manycore architecture that shares some features with a GPU.
- FPGA are doing an interesting entry in this world for throughput applications (streaming à la Hadoop). Google and Microsoft are investing quite a bit in FPGAs. Intel recently bought Altera to put FPGAs on the same die as their CPUs.
- DSPs are also trying to find a place in this market for solving specific problems with a special focus on energy efficiency.
- Some investigations are also underway to create ASICs to boost some specific operations (Tensor Processing Unit for the TensorFlow Framework of Google).

FPGAs and DSPs are far from offering general-purpose capabilities as of 2017. Yet some efforts exist to make them easier to use through ports of the OpenCL environment.

Concerning architecture integrating different type of heterogeneous nodes, these are motivated by the necessity to host in the same system, for the same user workload resources with different requirements. This is typical for many workloads implemented as workflows, where the users are asking to improve the performance of the whole process end-to-end and not of a single component. In approaching next generation systems (exascale) the number of these use cases are expected to grow, given the fact that pre-and post-processing could be easily HPC tasks themselves. Then, from an architecture point of view, there is the need to couple a scale-out throughput optimized partition of a system with a partition optimized to perform other tasks of the workload (often implemented through a high-level workflow, that could be run/managed by a high-level workflow manager, e.g. UNICORE).

Heterogeneous Intel based systems, featuring an Intel Xeon Phi partition together with a Xeon E5 partition integrated in the same Intel Omnipath or IB network are typical examples of this kind of architecture.

Moreover, one of the consequences of the fact that traditional (latency based) microprocessor functional units do not improve in terms of computational performance, at the same pace as in the past, is that research and innovation is focusing in specialized functional units, that are being integrated in larger microprocessors.

In this respect, we're seeing the growing importance of single and even half precision floating point computation engines. This impacts the architecture of the chips that will be available to scientists in the future. Yet evolution is not concentrated only on the compute part of the chips. A number of initiatives try to federate the different units present on a die with an interconnect what is adopted by many vendors. Here we can cite the CCI-X, GEN-Z or OpenCAPI initiatives. Their goal is to introduce heterogeneity at the lowest level possible (i.e. the die or the SOC) moving away from the model of the daughter card, a classical option for GPUs.

In general, as we are getting towards more powerful computers, we note that a "one size fits all" computer is no more possible to address the needs of a large and rich community of scientist coming from various domains. Heterogeneity will be the norm where, hooked to a shared node interconnect, the users will have a choice of partitions (be it ManyCore, GPUs or FPGAs) inside the supercomputer that will match the particularities of their codes. This trend exists already in Europe (Marconi, Tera1000) or worldwide (Trinity, Cori).

10 Conclusions

This second Technology Watch deliverable of PRACE-4IP Work Package 5 gives an updated overview of HPC and Big Data trends, in terms of technologies and with some market and business analysis hints. It is meant to complement the results of the first Technology Watch deliverable D5.1 by highlighting the new trends in HPC appeared within the last year, as well as giving more emphasis on Data Storage, Processing and Management (Section 5) and major paradigm shifts in HPC technologies that will most probably lead to new use cases and technologies for supercomputers in the near future (Section 9).

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 SC16 for this current report), other HPC events and public literature, direct (non-NDA) contacts with vendors and direct participation of WP5 members in a diversity of European projects or initiatives. Technical as well as operational aspects related to computing infrastructures are further investigated in Task 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 "Best Practices for HPC Systems Commissioning".