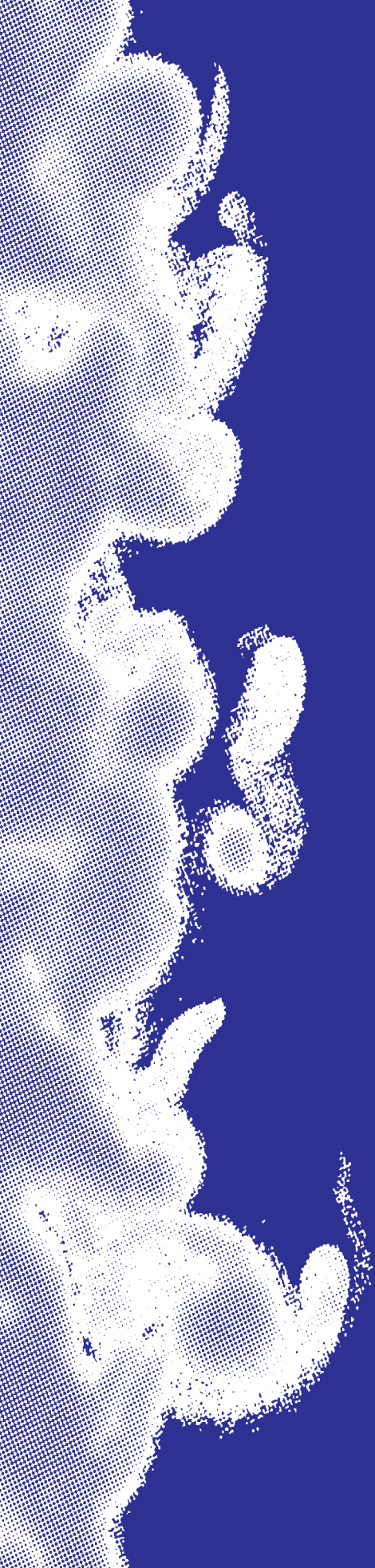


STRATEGIC RESEARCH AGENDA 2017

European Multi-annual
HPC Technology Roadmap

ETP 4 HPC

EUROPEAN TECHNOLOGY PLATFORM FOR HIGH PERFORMANCE COMPUTING

e dci

European Extreme Data & Computing Initiative

CONTENTS

A. GLOSSARY OF TERMS

ABFT	Application Based Fault Tolerance
API	Application Programming Interface
BD	Big Data
BDEC	Big Data and Extreme Computing, www.exascale.org
BoF	Birds-of-a-Feather
CAGR	Compound annual growth rate
CDN	Content Delivery Network
CERN	European Organization for Nuclear Research, www.home.cern
CFD	Computational Fluid Dynamics
CoE	Centres of Excellence in Computing Applications
DNS	Direct Numerical Simulation
DoE	Department of Energy
cPPP	Public-Private Partnership
EESI	European Exascale Software Initiative, www.eesi-project.eu
EOSC	European Open Science Cloud
EsD	Extreme-Scale Demonstrators
ETP4HPC	European Technology Platform for High Performance-Computing, www.etp4hpc.eu
EU	European Union (of 28 Member States)
EU+	European Union (28 Member States) plus Norway and Switzerland
EXDCI	European Extreme Data & Computing Initiative, www.exdci.eu
flops	floating point operations per second
FP7	Framework Programme 7, https://ec.europa.eu/research/fp7/index_en.cfm
GDP	Gross Domestic Product
GSS	Global System Science
HBM	High-Bandwidth Memory
HBP	Human Brain Project, www.humanbrainproject.eu
HiPEAC	High Performance and Embedded Architecture and Compilation, www.HiPEAC.net
HPC	High Performance Computing
HPDA	High-Performance Data Analysis
ICT	Information and Communication Technologies
IDC	International Data Corporation, www.idc.com
IESP	International Exascale Software Project
IPCC	Intergovernmental Panel on Climate Change
IPCEI HPC-BDA	an Important Project of Common European Interest in the area of HPC and Big Data, https://ec.europa.eu/commission/commissioners/2014-2019/oettinger/blog/luxembourg-launches-supercomputing-project_en
ISC	International Supercomputing Conference, www.isc-hpc.com
ISV	Independent Software Provider
ITER	("The Way" in Latin) fusion energy research collaboration, www.iter.org
L2ERTM	Least Squares/Residual Shot Elastic RTM
LES	Large Eddy Simulations
NVRAM	Non-volatile random-access memory
PCIe	PCI Express (Peripheral Component Interconnect)
PCP	Pre-Commercial Procurement
PPI	Public Procurement of Innovative solutions
PRACE	Partnership for Advanced Computing in Europe, www.prace-ri.eu
QoS	Quality of Service
ROI	Return on Investment
SC	Supercomputing Conference, www.supercomputing.org
SHAPE	SME HPC Adoption Programme in Europe
SKA	Square Kilometre Array, www.skatelescope.org/project/
SME	Small and Medium-size Enterprise(s)
SRA	Strategic Research Agenda, www.etp4hpc.eu/sra
SWOT	Strengths, Weaknesses, Opportunities, Threats
TCO	Total Cost of Ownership
Top500	the ranking of top world supercomputers at www.top500.org
TTI RTM	Tilted Transversely Isotropic Reverse Time Migration
WP	Work Programme

B. FOREWORD

This is our third Strategic Research Agenda (SRA 3) and the second fully revised edition (following the first full edition / SRA 1 / in 2013 and an incremental update / SRA 2 / in 2015). We maintain an up-to-date European HPC technology roadmap, validated by the European HPC ecosystem, as one of the key deliverables of our association. As usual, the process of writing the SRA was an open one, with all the members of ETP4HPC having an opportunity to contribute their expertise. This is the result of the collective work of nearly two hundred and thirty experts in eight technical working groups as well as some other non-technical task forces and technical and non-technical experts. I would like to thank our members for their continuous support and involvement.

There is an emphasis in this SRA on the concept of European HPC system prototypes, which we call 'Extreme-Scale Demonstrators' (EsDs). The other focal area included is that of Big Data, which has emerged as an increasingly pervasive topic and a priority since our last SRA. Both topics received a lot of attention in the course of developing the material for this SRA. We discussed the EsDs with all the stakeholders involved and we maintain a dialogue with the European Big Data Value Association (BDVA) to align our priorities.

The first challenge ahead of us is that of producing EsDs using the technology developed based on our SRA. This will be a test of our work – the contents of our SRAs has served as the basis for project definition and assessment for the entire Horizon 2020 programme. The second challenge is to bridge the gap between HPC and other areas, such as Big Data. If it is all going to be one and the same thing, how will it work?

I strongly believe this research agenda will strengthen the European HPC ecosystem and contribute to Europe's academic and industrial competitiveness.

Jean-Pierre Panziera
ETP4HPC Chairman
November 2017

Figure 1 .
The ETP4HPC Steering Board and SRA Experts — May 2017.

C. EXECUTIVE SUMMARY

This Strategic Research Agenda (SRA) outlines the European research priorities in the area of HPC technology. The provision of HPC technology is viewed here as one of the three pillars of European HPC, alongside the infrastructure and application expertise. The importance of HPC is now widely recognised and European HPC has a good momentum due to the investments made and the programmes launched. Europe is improving its position to compete with other regions in the area of HPC technology provision. The level of HPC deployment in Europe matches its worldwide economic and academic position. The development of European HPC technology should continue along the seven main research lines (HPC System Architecture and Components, System Software and Management, Programming Environment, Energy and Resiliency, Balance Compute, I/O and Storage Performance, Big Data and HPC usage Models, Mathematics and algorithms for extreme scale HPC systems) but it also includes the concept of prototyping ('Extreme-Scale Demonstrators') to test the readiness of the European technology projects, vendors and users to produce a globally competitive HPC system. The area of Big Data has received a special attention and work will continue to synchronise HPC technology and Big Data solutions. We also believe that two other areas, namely support mechanisms for SMEs and Start-ups and Education and Training needs to remain as priorities to help Europe compete in this complex market.

D.
HOW TO USE THIS DOCUMENT

Chapter 1 introduces the objectives of this document and its context. Chapter 2 presents 'the case for European HPC technology'. It also summarises the current European HPC strategy in a short review of the current tools and mechanisms in place which serve to develop European HPC. Chapter 3 contains a review of the European HPC ecosystem and gives a short overview on the status of HPC in other geographies.

Chapter 4 explains the structure of the technical Chapters (5 to 8) and contains a detailed description of the requirements for future research priorities meeting the needs of scientific and industrial users. This is followed by the technical (5 – trends, 6 – research priorities, 7 – milestones and 8 – Extreme-Scale Demonstrators) and non-technical (9) areas of this roadmap. Section 10 provides the conclusions.

Any general source references used the introductory part of text are included at the end of the document. Any direct technical references used in the technical part (Section 4 onwards) are listed as footnotes throughout the text.

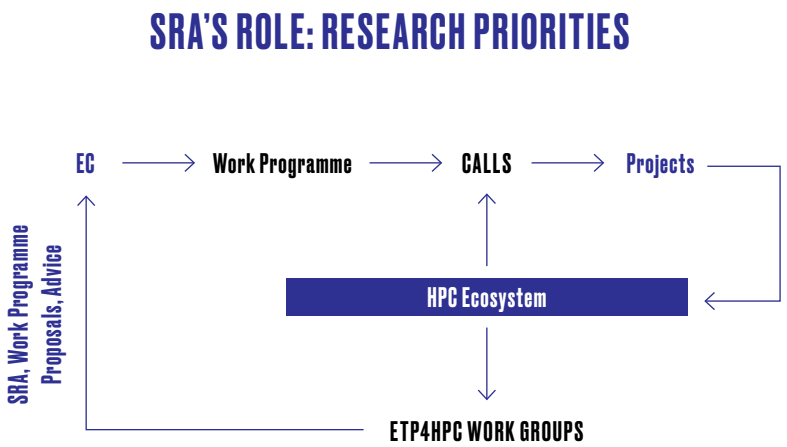
1.

INTRODUCTION AND CONTEXT

SRA 3 is the second fully re-worked edition of ETP4HPC's Strategic Research Agenda, following its first issue in 2013 and its incremental update in 2015. The role of SRA 3 is to define Europe's roadmap towards HPC technology provision[1] at Exa-scale and beyond (and, as a consequence, throughout the broader ICT landscape). The key motivation of ETP4HPC is to increase the global market share of the HPC technology developed in Europe. This means that future systems produced in Europe need to be able to compete with systems from other geographies. These systems need to meet the requirements of European (and global) scientific and industrial users and facilitate the pervasive use of HPC.

In this process, ETP4HPC, as part of the European HPC Ecosystem, issues its SRA, which is then used by the EC as a recommendation in formulating its research programme. In the next step, the EC's open calls for proposals are announced and the guidelines included in this SRA 3 are expected to be used as a reference for the call objectives and assessment criteria for project proposals submitted to the EC's HPC technology Work Programme (WP) 2018-2020. Any project proposal submitted within the Programme following the issue of the SRA should address its research recommendations, defined as research priorities and research milestones. In particular, as SRA3 is being issued in Q4 2017, any project funded under WP18-20 (the last part of the EC's Horizon 2020 Framework Programme) should cover milestones included in this document.

Figure 2.
The SRA influences the process of defining EC calls for proposals in the area of HPC technology. The guidelines included in the SRA are expected to be used as a reference for the call objectives and assessment criteria for project proposals submitted.



[1] The term 'technology' is used here to denote the entire HPC system stack as per the ETP4HPC HPC Model (Section 4.1 – including services and solutions).

Throughout the H2020 timeline, there is an updated, validated by the European HPC ecosystem SRA in place to serve as a reference for the projects participating in the EC HPC research programme. The FET-HPC[2] part of the programme concerns the development of basic HPC technology. The CoE[3] sub-programme supports Centres of Excellence in Computing Applications, consolidating the European HPC application expertise. The ecosystem is supported by a series of Coordination and Support Actions[4], which orchestrate the European HPC strategy. It is important to note that that some relevant elements of the European HPC effort might fall into other (non-HPC) programme parts such as LEIT[5].

[2] Future and Emerging Technologies – High-Performance Computing

[3] Centres of Excellence (in Computing Applications)

[4] CSA

[5] Leadership in Enabling and Industrial Technologies

HORIZON 2020 TIMELINE

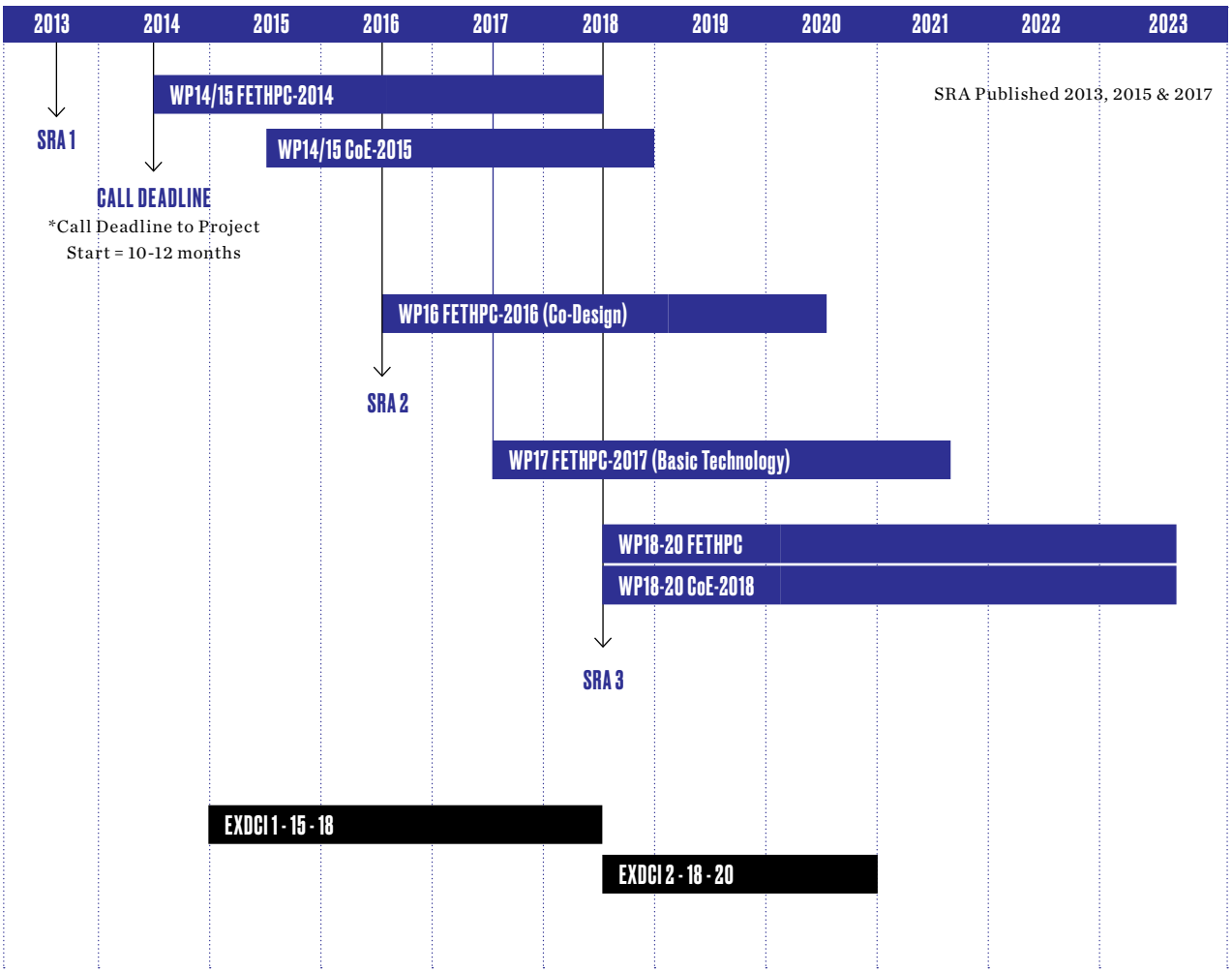
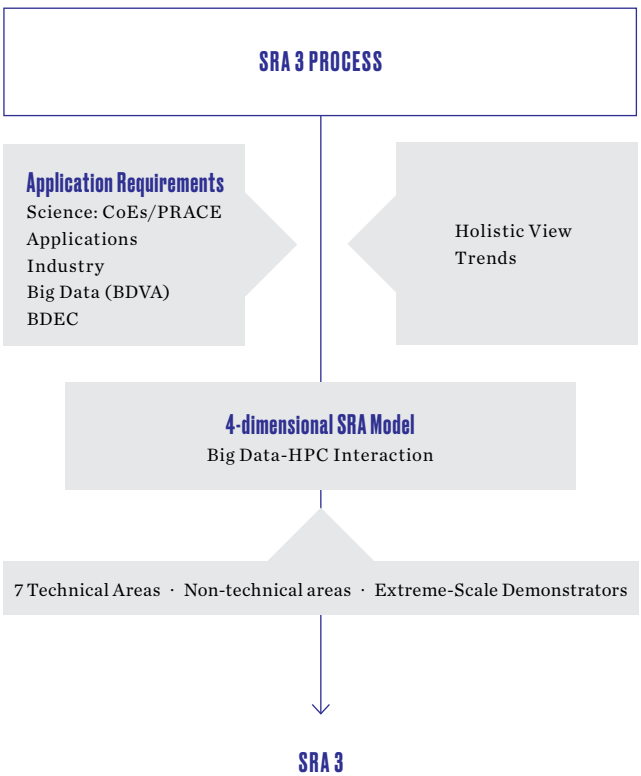


Figure 3.
The timings of the SRA and various Horizon 2020 EC HPC technology work programme parts - there is a valid SRA in place at any time within the programme.

1.1
THE PROCESS OF PREPARING SRA 3

The process of preparing SRA 3 began in March 2017 with the definition of the steps required. .

Figure 4.
The process of preparing SRA 3, including the inputs used, the analysis undertaken and the outputs.



First, the requirements of applications were identified. In this step, ETP4HPC collected input from the following areas (their input is included in Section 4.3 – Application Requirements):

- Scientific Applications: Centres of Excellence in Computing Applications (CoEs[6]) and PRACE[7] Application Scientific and industrial users (through the 'Scientific Case 2017' [PRACE])
- Industrial users
- BDEC's recommendations[8]
- HiPEAC[9]'s vision.

A special emphasis was placed on the needs of the European Big Data Community represented by BDVA[10].

Then, an analysis of the current state of European HPC technology was carried out (Section 3). In this part, a holistic view of it and some recommendations are included. The results of these analyses are fed into the following eight working areas, in line with the previous issue of the SRA:

- HPC System Architecture and Components
- System Software and Management
- Programming Environment
- Energy and Resiliency
- Balance Compute, I/O and Storage Performance
- Big Data and HPC Usage Models
- Mathematics and Algorithms for extreme scale HPC systems
- Extreme-Scale Demonstrators
- And also non-Technical Areas: Education and Training, SMEs and Start-ups

[6] https://ec.europa.eu/programmes/horizon2020/en/news/overview-eu-funded-centres-excellence-computing-applications
[7] Partnership for Advance Computing in Europe, www.prace-ri.eu
[8] Big Data and Extreme Computing, www.exascale.org
[9] High Performance and Embedded Architecture and Compilation, https://www.HiPEAC.net/
[10] Big Data Value Association, www.bdva.eu

All ETP4HPC members were invited to participate in technical working groups (WGs), mirroring the categories above and led by selected ETP4HPC member organisations. The task of these working groups was to define the research priorities (Section 6) and milestones in the corresponding areas (Section 7). This work took place through conference calls within the groups and a workshop involving all working group leaders. All ETP4HPC members had been able to review and comment on the last draft of this document before it reached the approval stage at the ETP4HPC Steering Board level.

Within this process, a special emphasis was placed on the concept of the Extreme-Scale Demonstrators – prototypes of European exascale supercomputers (EsDs – Section 8). To facilitate the definition of the scope of these systems, a total of five[11] workshops have been held in 2016 and 17, with two of these events falling within the period of SRA 3 preparation:

- A 'Round-table' workshop aiming to present the potential contributions of the European HPC technology projects, involving also the Centres of Excellence in Computing Applications (CoEs), system integrators and HPC Centres (May 2017, during the EXDCI European HPC Summit Week)

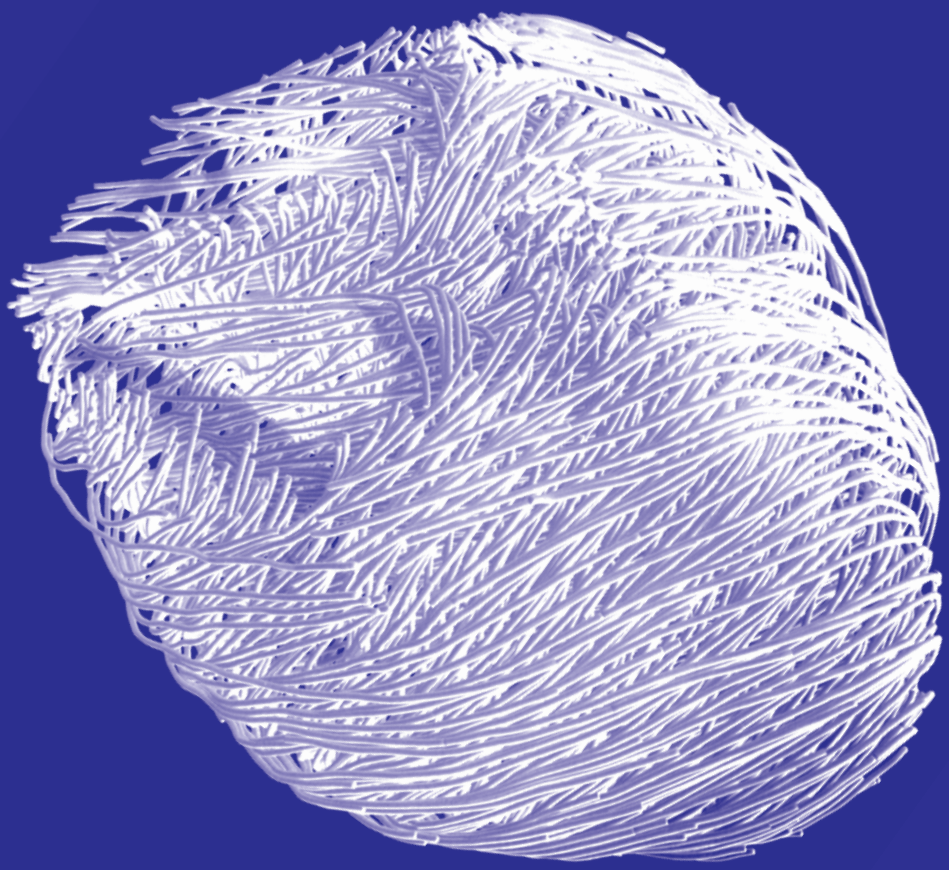
- A workshop dedicated to the potential industrial use of the EsDs, involving industrial users of HPC and Independent Software Vendors (ISVs). The invited parties were requested to present their application domains and describe their interest in participating in future EsD projects.

The second area of focus in this SRA is Big Data and its HPC system requirements. In order to synchronise the technology of the two areas, a workshop involving ETP4HPC and BDVA (Big Data Value Association) was held in July 2017, within the period of preparation of this SRA[12].

[11] These five EsD-related workshops have provided an opportunity for the following entities to express their comments on the concept of the EsDs and their potential contributions to them: 1/all FETHPC and other technology projects, CoE projects and system integrators, 2/all CoE projects and other application users, 3/ system integrators, and (within the scope of SRA 3) 4/ all FETHPC and other technology projects and 5/industrial HPC users and ISVs.

[12] Another Big Data-related workshop was held in Sept 2016 with an objective to analyse selected Big Data use areas where HPC could contribute.

2.



THE CASE FOR EUROPEAN HPC TECHNOLOGY

Throughout the period in which ETP4HPC has been in existence, the value of HPC for European science and for its economy and society has been demonstrated repeatedly [as confirmed by IDC 1]. European has developed its own vibrant HPC ecosystem (Section 3) which has contributed to the growth of the economy, science and the resolution of Europe's Grand Societal Challenges.

HPC is now widely recognised as an indispensable tool by users in academia, industry (manufacturing and services) and also by decision makers. A significant progress has been made to consolidate and develop the European HPC resources in terms of HPC infrastructure, technology provision and application expertise, which has been supported by various European-level policy mechanisms: one of the main goals of the European Horizon 2020 HPC programme is an efficient and well-coordinated HPC ecosystem in Europe.

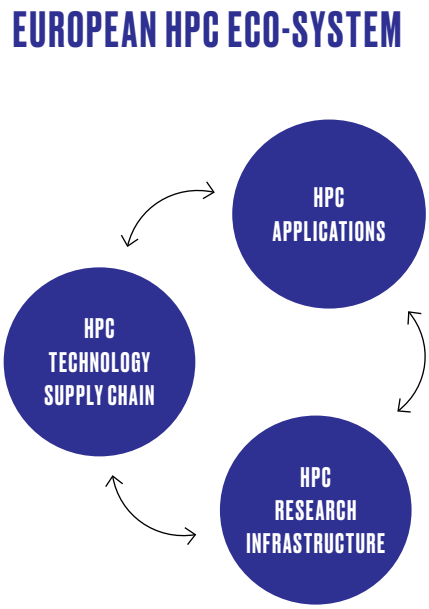


Figure 5.
The European HPC Ecosystem and its positive impact on the European economy, science and society.

2.1
THE VALUE OF HPC

2.1.1
HPC as a Scientific Tool

Scientists from throughout Europe increasingly rely on HPC resources to carry out advanced research in nearly all disciplines. European scientists play a vital role in HPC-enabled scientific endeavours of global importance, including, for example, CERN (European Organisation for Nuclear Research), IPCC (Intergovernmental Panel on Climate Change), ITER (fusion energy research collaboration), and the newer Square Kilometre Array (SKA) initiative. The PRACE Scientific Case for HPC in Europe 2012 – 2020 [PRACE] lists the important scientific fields where progress is impossible without the use of HPC.

2.1.2
HPC's Contribution to the Economy

There are a number of reports [IDC[13]1-3, HPC User Forum, Ezell and Atkinson], on the Return on Investment produced by applying HPC in an industrial environment[14]. Although they differ in the ratios established, they all conclude that:

- HPC produces returns faster and higher than most other technologies
- Most companies using HPC find it indispensable for their competitiveness

The most important sectors of the European industry rely on HPC: automotive, aviation, energy, oil and gas, pharmaceutical, etc. Industry accounted for 24.5% of EU GDP in 2015 [Statista]. Studies by IDC and others firmly established the link between HPC and industrial competitiveness. Reports confirm the immense contribution of HPC in job and GDP creation in, for example, the oil and gas industry, healthcare and manufacturing, for example. Manufacturing contributed about 30 million jobs and 16% of EU GDP (€6,500 billion) in 2013 [IDC 1], and the European Commission aims to increase that figure to 20% by 2020. HPC enables smart manufacturing that could create new manufacturing jobs and return some lost manufacturing jobs to Europe. The region spent about €450 million on the HPC ecosystem for manufacturing in 2013 and will spend about €638 million in this sector in 2018. A substantial, growing portion of this spending is by manufacturing SMEs, who, like larger manufacturing firms, employ HPC to accelerate innovation[15].

A strong European HPC Value Chain can also strengthen the European economy by holding a share of the global market and e.g. through job creation and the provision of novel technologies (to be used in other areas).

2.1.3
HPC's Contribution to Society

HPC is increasingly important in the addressing of the Grand Social Challenges, the key issues facing our region [EC 1-2].

- Health, demographic change and wellbeing (e.g. personalised medicine)
- Food security, sustainable agriculture and forestry, marine and maritime and inland water research, and the Bioeconomy (e.g. simulations of sustainability factors /e.g. weather/)
- Secure, clean and efficient energy (e.g. fusion energy)
- Smart, green and integrated transport (e.g. performance management)
- Climate action, environment, resource efficiency and raw materials (e.g. gas and oil search)
- Europe in a changing world - inclusive, innovative and reflective societies (e.g. smart cities)
- Secure societies - protecting freedom and security of Europe and its citizens (e.g. through data analysis).

For example, HPC is applied in advanced medical research, biomedicine, bioinformatics, epidemiology, and personalised medicine—including «Big Data» aspects, e.g. in the improvement of cancer treatments. As an example, Europe spent about €173 million on the HPC ecosystem for weather/climate in 2013 and will spend about €230 million in this sector in 2018 - the provision of meteorological services to agriculture, industry

[13] https://www.idc.com/

[14] IDC field research confirms that European HPC investments are producing excellent returns-on-investment (ROI) for industry. Each euro invested in HPC on average returned €867 in increased revenue/income and €69 in profits [IDC 1-2]. The total increased revenue for the 59 HPC-enabled, quantifiable projects was €133.1 billion, or about €230 million per project on average. In 2015, a DoE-funded study [HPC User Forum] meticulously analysed several hundred concrete projects in the United States. The study revealed an approximate profit of $43 per dollar that was invested in HPC. The average number of years before enterprises realize a return on their HPC investments is approximately three years [IDC 2]. In a global IDC study, 97% of companies that had adopted HPC said they could no longer compete or survive without it [IDC 3].

and society in general is impossible without HPC. Similarly, understanding climate change and the impact of climate change is only possible on the basis of HPC-enabled research.

2.2
THE CURRENT EUROPEAN HPC STRATEGY

In this chapter, we focus on the HPC strategy as defined in various communications by the European Commission. Several member states (e.g. France) have separate national HPC strategies, which will not be discussed in this SRA. Currently, the open collaboration between the EC and member states in order to provide a legal framework coordinating the national and European efforts (see Chapter 2.2.3 on EuroHPC below) has a good momentum which should be maintained.

The overall strategic goal outlined recently by the EC is to develop a thriving European HPC ecosystem with three areas of emphasis:

· Infrastructure: Put in place the capacity of acquiring leadership-class HPC systems

· Technology: Securing an independent European HPC systems and technology supply

· Applications: Excellence in HPC applications and widening the use of HPC

2.2.1 The European Cloud Initiative

The foundations of this European HPC strategy were laid by the *EC Communication High-Performance Computing: Europe's place in a Global Race of 2012*[16] [EC 3]. In 2015, a policy called "Single Digital Market"[17] was announced, followed in 2016 by the announcement of the "European Cloud Initiative - Building a competitive data and knowledge economy in Europe"[18]. This initiative [EC 3, 4] aims to strengthen Europe's position in data-driven innovation, improve its competitiveness and cohesion, and help create a Digital Single Market in Europe. It aims to provide European science, industry and public authorities with:

· A world-class data infrastructure to store and manage data

· High-speed connectivity to transport data

· More powerful High-Performance Computers to process data

This initiative has three central objectives:

· Establishing and maintaining the European Open Science Cloud

· Establishing and maintaining the European Date Infrastructure

· Widening access to, and trust in, the above

According to its charter, the European Cloud Initiative calls for the support of EU Member States to develop a High-Performance Computing ecosystem based on European technology, including low-power HPC chips. The goal is to have Exascale supercomputers based on European technology in the global top three by 2022.

2.2.2 IPCEI[19]

The IPCEI HPC-BDA (an Important Project of Common European Interest in the area of HPC and Big Data Applications) is an initiative led by Luxembourg, France, Italy and Spain, which should result in an EC support instrument to enable a streamlining of investments in the area of HPC systems. It has the following objectives:

· To ensure European industrial expertise in key HPC technologies (placing an emphasis on safety and security)

· To support and create new uses of HPC by the industry and develop industrial applications that require exascale supercomputing and data infrastructures

· To guarantee access to world-class HPC facilities for public and private research

· To correct market failures and fragmentation at both supply and application levels

[15] Further data [EC 2] reports the following contributions of the sectors of the European economy: manufacturing: 6,500B Euro of GDP and 30 million jobs, oil and gas – 440B Euro/170K jobs, pharma – 800B Euro/40% of the EU worldwide marker shares for medicine (1,000B Euro public spending /10% of the EU's GDP/).

[16] http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2012:0045:FIN:EN:PDF

[17] https://ec.europa.eu/digital-single-market/en/policies/shaping-digital-single-market

[18] https://ec.europa.eu/digital-single-market/en/news/communication-european-cloud-initiative-building-competitive-data-and-knowledge-economy-europe

[19] https://ec.europa.eu/commission/commissioners/2014-2019/oettinger/blog/luxembourg-launches-supercomputing-project_en

2.2.3
EuroHPC[20-21] — Europe as a global player in High Performance Computing

In March 2017, ministers from seven European countries (France, Germany, Italy, Luxembourg, Netherlands, Portugal and Spain[22]) signed a declaration to support the next generation of computing and data infrastructures, a European project of the size of Airbus in the 1990s and of Galileo in the 2000s. The Member States plan to establish EuroHPC for acquiring and deploying an integrated world-class high-performance computing infrastructure capable of at least 10^{18} floating point calculations per second (exascale computers). This will be available across the EU for scientific communities, industry and the public sector, no matter where the users are located. Goals are to deploy two pre-Exascale systems by 2019/2020 and two Exascale systems by 2022/2023 of which at least one will be based on European technology. Also, EuroHPC aims at developing test-beds applications in science, industry and public administrations.

A governance structure for EuroHPC is expected by the end of 2017, detailing the interaction, roles and responsibilities of the players (EC, joining member states and industrial partners).

IPCEI-HPC-BDA and EuroHPC need to be seen as independent initiatives with partially overlapping goals.

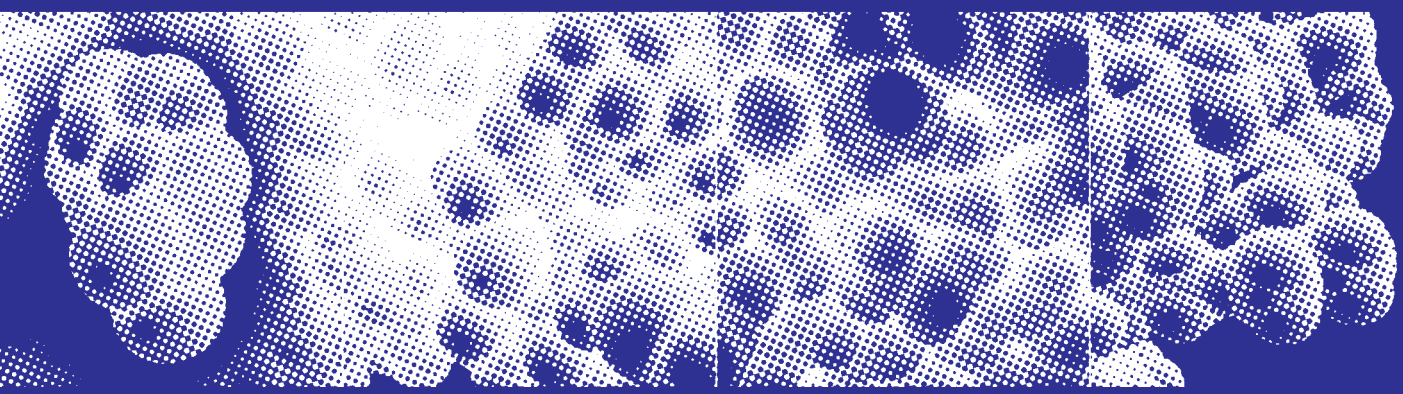
[20] **https://ec.europa.eu/digital-single-market/en/news/eu-ministers-commit-digitising-europe-high-performance-computing-power** - please refer to this page for further information.

[21] **https://ec.europa.eu/digital-single-market/en/news/eurohpc-initiative-speeds-its-pace**

[22] Other countries have joined since the initial declaration (e.g. Belgium, Slovenia, Bulgaria, Switzerland, Greece and Croatia).


M-ALG-9
for vertical integration of algorithms
ished together with the validation of
r, ease of implementation, tuning and
optimisation.
2018

3.



THE EUROPEAN HPC ECOSYSTEM – WHERE WE ARE AND WHERE WE SHOULD GO

The purpose of this chapter is to analyse the current position of, and lay out a future strategy for, of the European HPC landscape and Technology Value Chain, with an objective to increase the global market share of European HPC technology. Its conclusions – i.e. the strategic priorities of European HPC technology provision – will be reflected in the design and content of the technical part of this document.

First, the strategic position of the European HPC environment and related elements are presented (including the strategic position of European HPC technology provision). Then, a holistic view of the European HPC ecosystem is presented as developed by the EXDCI project. This is followed by an analysis of the performance of other regional HPC technology ecosystems, with which Europe will have to compete.

Analyses available and other sources:

IDC's and Hyperion's analysis of European HPC [IDC 1, Hyperion – a report commissioned by ETP4HPC] is a source of data and strategic directions for the entire European HPC ecosystem. However, it does not focus on the HPC technology and its growth (it rather focuses on general HPC policies and instruments). IDC is also a source of other market-related data. Other available roadmaps (e.g. BDVA, HiPEAC and ReThinkBig) cover areas that interact with our community.

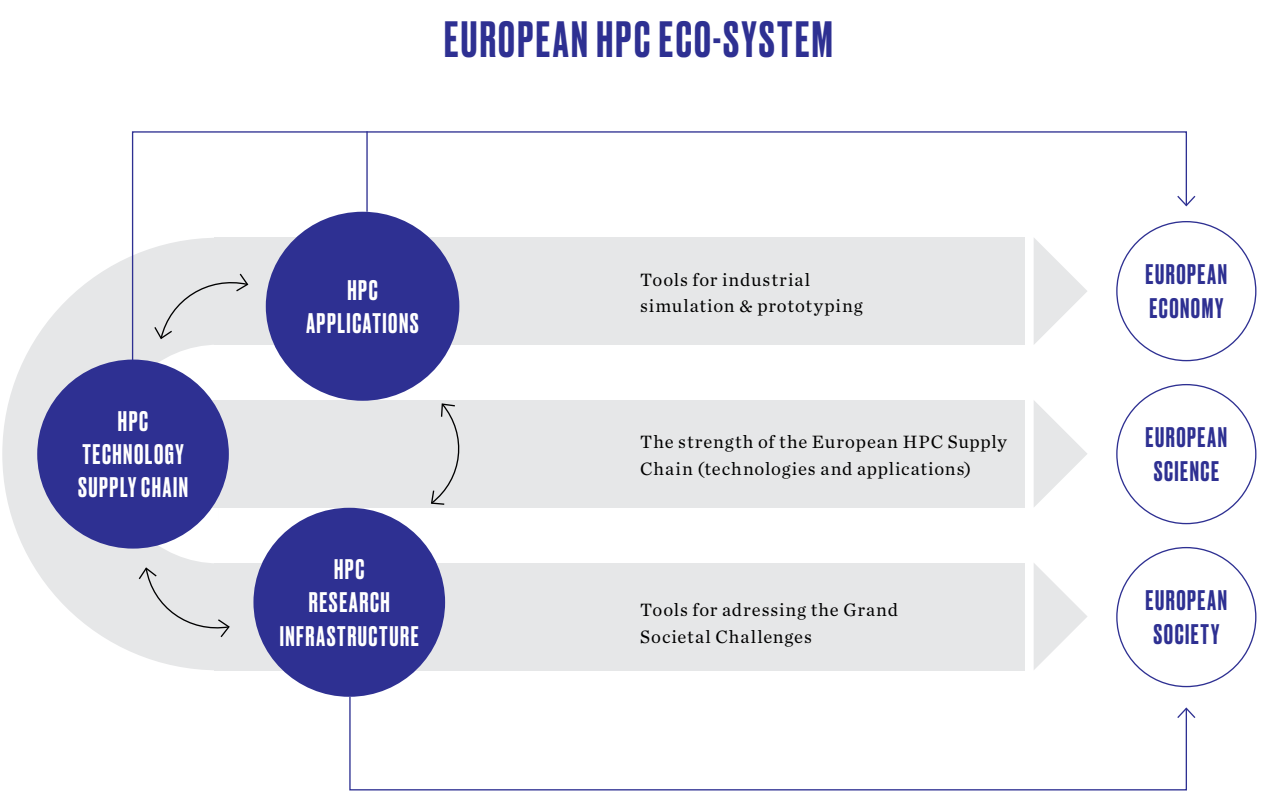
3.1
THE EVOLUTION OF THE EUROPEAN HPC ECOSYSTEM (2012-17)

The following section analyses the changes in the European HPC technology since the establishment of ETP4HPC.

The main achievement of European HPC since the issue of SRA 1 in 2013 is the establishment of a vibrant HPC Ecosystem that covers all the aspects of HPC access, use and provision and is also able to collaborate effectively with other stakeholders and ecosystems. Below follows a summary of its elements.

The European HPC ecosystem as defined in the 2012 EU publication *High-Performance Computing: Europe's place in a Global Race* [EC 3] consists of three pillars: Technology, Infrastructure and Application Expertise.

Figure 6.
The European HPC Ecosystem and its impact on the European economy, science and society - the Technology and Application pillars are the EC's partners in the HPC cPPP.



3.1.1 European HPC Technology

3.1.1.1 European HPC Technology Suppliers

The latest reports [Hyperion, IDC 1,4] confirm the US and Japan supremacy in the area of HPC system provision. The largest European provider (Bull/Atos) held less than 4% in the global 2015 HPC Server market. The share of the same provider in the European market oscillates between 3% and 6% between 2014 and 2016. The global HPC server market is supposed to grow by 32.3% from 2016 to 2021, while the broader market (with the addition of middleware, applications and services) is meant to grow at a CAGR[23] of 6.2%, or by 34.8%. The European HPC server market is supposed to grow at a CAGR rate of 6.2% or 35% in total in the same period. Ownership changes might further affect the European HPC supplier scene.

A new Hyperion Research study **"The Status and Prospects of European Suppliers of High Performance Computing (HPC) Products and Services"** conducted for the European Extreme Data & Computing Initiative (EXDCI) in 2017, substantially augments the data and analysis provided previously on the anticipated growth of European HPC technology providers in the broader European HPC market (servers, storage, software, networks). Indigenous European suppliers will grow their share from 6.4 % in 2016 to 11.3% in 2021 (from 337 million Euro in 2016 to 839 million Euro in 2021). The share of International suppliers with R&D in Europe decreases from 77% to 72% in the same timeframe and the share of suppliers with no R&D in Europe stays flat at 17%. This projection shows two sides of the same coin: an almost doubling share for indigenous European suppliers, but at an overall low percentage level. This interpretation reinforces the requirement for effective methods to strengthen the position and capabilities of indigenous European suppliers, most of them being SMEs.

3.1.1.2 ETP4HPC

ETP4HPC represents the stakeholders of the European HPC Technology Value Chain. ETP4HPC is an association that facilitates interaction between its members and represents their interests in developing the strategic R&D&I HPC strategy. The ETP4HPC SRA shapes the future HPC technology landscape in Europe. ETP4HPC is also the partner in a contractual Public-Private Partnership with the EC.

3.1.1.3 FETHPC Projects

As a result of the EC H2020-FETHPC-2014[24] call, there are currently 19 HPC technology research projects running in Europe and two more co-design projects are starting at the time of the writing this document that have arisen from the subsequent FETHPC-01-2016 call[25] (DEEP-EST[26] and EuroExa[27]). Their objectives are aligned with the goals and milestones of SRA-1 (the 2014 call) and -2 (the 2016 call). Further projects (in the area of basic technology and co-design) will operate following the closure of FETHPC-02-2017[28] and further proposed calls. As the project selection process is open and competitive, the projects demonstrate a significant degree of technological diversity. However, some areas (e.g. System Software) are not adequately covered at the moment. These and subsequent projects are expected to feed their technological results into the future European systems (cf. the concept of Extreme-Scale Demonstrators – Section 8).

[23] CAGR = Compound Annual Growth Rate

[24] https://ec.europa.eu/research/participants/portal/desktop/en/opportunities/h2020/calls/h2020-fethpc-2014.html#c,topics=callIdentifier/t/H2020-FETHPC-2014/1/1/1/default-group&callStatus/t/Forthcoming/1/1/0/default-group&callStatus/t/Open/1/1/0/default-group&callStatus/t/Closed/1/1/0/default-group&+identifier/desc

[25] http://ec.europa.eu/research/participants/portal/desktop/en/opportunities/h2020/topics/fethpc-01-2016.html

[26] http://www.deep-projects.eu/

[27] https://twitter.com/euroexa

[28] http://ec.europa.eu/research/participants/portal/desktop/en/opportunities/h2020/topics/fethpc-02-2017.html

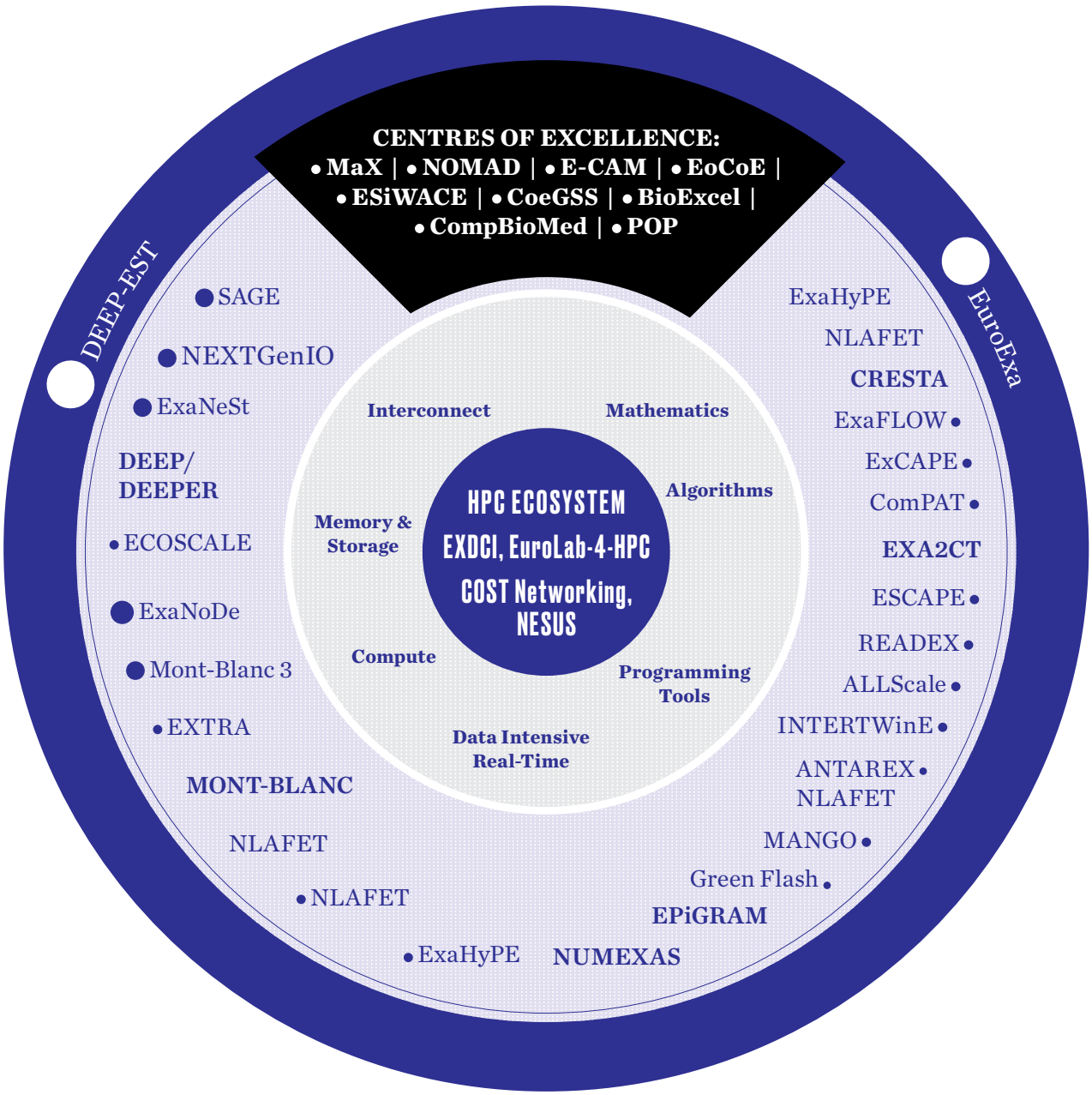


Figure 7.
The currently running FETHPC technology (inside circles), application (top) and co-design (outer circle) projects. The size of the blue circles indicates the funding committed. E.g. SAGE = 7.9 million Euro.

3.1.2
European HPC Infrastructure – PRACE[29]

Europe's overall HPC capabilities have made impressive progress in recent years [IDC 1]. Largely through the PRACE Research Infrastructure, Europe has narrowed the former wide gap separating the most capable U.S. and Japanese supercomputers from their European counterparts. The aggregate peak performance of the Europe-based supercomputers rose more than ten-fold from 4.3 Petaflops in November 2010 to 48.9 Petaflops four years later. (During this same period, the aggregate peak performance of all top 50 supercomputers grew by a lesser 7.6 times, from 32.8 Petaflops to 249.7 Petaflops. Europe's share of these totals increased from 13.1% in November 2010 to 19.9% in November 2014).

The mission of PRACE (Partnership for Advanced Computing in Europe) is to enable high impact scientific discovery and engineering research and development across all disciplines to enhance European competitiveness for the benefit of society. PRACE seeks to realise this mission by offering world class computing and data management resources and services to researchers from academia and industry through a centralised peer review process.

The computer systems and their operations accessible through PRACE are provided by PRACE members. Four hosting members (BSC[30] representing Spain, CINECA[31] representing Italy, GCS[32] representing Germany and GENCI[33] representing France) secured funding for the initial period from 2010 to 2016. In 2017, PRACE has engaged in the second period of the Partnership, securing the operations of the infrastructure until 2020, and including a fifth Hosting Member, CSCS[34] representing Switzerland. During this second phase, PRACE will offer an initial performance close to 70 Petaflops in 7 complementary leading edge systems, offering a total of 4.000 million core hours per year (75 million node hours).

The European Scientific HPC Users are the users of the PRACE Infrastructure. EXDCI, the project tasked with coordinating the European HPC strategy, helps these users to define their needs in terms of HPC system specifications (see Section 3.1.5).

3.1.3
European HPC Application Expertise - CoEs

The European Centres of Excellence in Computing Applications (CoEs, a result of the EINFRA-5-2015[35] call by the EC[36]) consolidate the European HPC Application knowledge in eight domains and one transversal area. There is an established interaction channel between ETP4HPC's roadmap and the CoEs – they have been contributing to our SRA and related documents, in particular in the area of the Extreme-Scale Demonstrators (Section 8) and the Application Requirements, Section 4.3.1). Also, the CoEs take part in the cPPP discussions.

3.1.4
Contractual Public-Private Partnership for HPC

ETP4HPC is the European Commission's partner in the contractual Public-Private Partnership[37] (cPPP) for High-Performance Computing. The objectives of this cPPP are to:

- Develop the next generation of HPC technologies, applications and systems towards exa-scale

- Achieve excellence in HPC applications delivery and use

The HPC cPPP brings together technology providers and users via the ETP4HPC Association and Centres of Excellence (CoE) for computing applications. The cPPP focuses on technologies and applications (two of out of the three) pillars of the European HPC strategy along with training, education and skills development. It constitutes a forum for a detailed dialogue on the implementation of the European HPC strategy.

[29] www.prace-ri.eu

[30] Barcelona Supercomputing Center, www.bsc.es

[31] Consorzio Interuniversitario del Nord est Italiano Per il Calcolo Automatico, www.cineca.it/en

[32] The Gauss Centre for Supercomputing, www.gauss-centre.eu/

[33] Grand Équipement National de Calcul Intensif, www.genci.fr

[34] Centro Svizzero di Calcolo Scientifico, the Swiss National Supercomputing Centre, www.cscs.ch/

[35] http://ec.europa.eu/research/participants/portal/desktop/en/opportunities/h2020/topics/einfra-5-2015.html

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

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

3.1.5 European eXtreme Data and Computing Initiative (EXDCI[38])

The operation of the European HPC Ecosystem is facilitated by EXDCI – the European Extreme Data and Computing Initiative. This project provides tools and funding for the implementation of the most important ecosystem activities:

· The issue of the ETP4HPC SRA(s)

· The issue of the PRACE Scientific Case

· Analysing cross-cutting ecosystem issues

· KPI measurement

· International collaboration

· Education and Training

· Dissemination

The existence of this action ensures that the individual ecosystem elements do not operate in isolation and a high-level of synchronisation takes place among them.

3.1.6 EuroLab4HPC[39]

EuroLab-4-HPC is a two-year Horizon 2020 funded project with the commitment to build the foundation for a European Research Centre of Excellence in High-Performance Computing (HPC) Systems. It is a Coordination and Support Action (CSA), funded in the same call as EXDCI. It is coordinated by Chalmers University of Technology and it involves thirteen prominent research organisations across nine countries with some of the best HPC research teams in Europe. The project's aim is to boost European research excellence on the key challenges towards the next generations of high-performance computing systems, in order to build connected and sustainable European HPC leadership.

The project's main objectives are (1) to join HPC system research groups around a long-term HPC research agenda by forming an HPC research roadmap and joining forces behind it, (2) to define an HPC curriculum in HPC technologies and best-practice education/training methods to foster future European technology leaders, (3) to accelerate commercial uptake of new HPC technologies, (4) to build an HPC ecosystem with researchers and other stakeholders, e.g., HPC system providers and venture capital, and (5) to form a business model and organisation for the EuroLab-4-HPC excellence centre in HPC systems.

The Eurolab-4-HPC Long-Term Vision on High-Performance Computing was issued in August 2017. As an academic research vision for HPC, which concentrates on technical matters from the perspective of academic research, it complements the industry-led ETP4HPC SRA. The Eurolab-4-HPC Vision has a substantially longer-term time window, covering the post-exascale era of 2023–2030. The vision is structured around the topics of application requirements, convergence with data centres and cloud computing, disruptive hardware technologies, vertical challenges (including green ICT and resiliency), and system software and programming environment. ETP4HPC and Eurolab4HPC have cooperated to issue consistent and complementary roadmaps[40].

[38] https://exdci.eu

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

[40] This has primarily been by sending representatives to each other's meetings: Theo Ungerer (EuroLab-4-HPC) presented at the EXDCI Technical Meeting in Barcelona (21–22 September 2016) and the ETP4HPC SRA3 kickoff in Munich (20 March 2017). François Bodin (EXDCI) presented at the EuroLab-4-HPC Expert Working Group meeting in Lausanne (29–30 May 2017) and he reviewed the preliminary (2016) and final (2017) versions of the EuroLab-4-HPC Vision.

3.2
A HOLISTIC VIEW OF EUROPEAN HPC

The Petascale to Exascale transition is very complex, especially because it is not happening in isolation. At the same time, a data deluge is taking place. An Exascale definition limited to producing a machine capable of a rate of 10^{18} flops is of interest to only a few scientific domains. The main issue is to deal with the data generated by sensors as well as the numerical simulations themselves. While in the European Exa-scale Software Initiative[41] (EESI) it was clear that the data issue would be of crucial importance, during the EXDCI time-frame the focus has shifted to the convergence of extreme data and computing with new considerations such as Edge computing, in-transit computing, etc. This new focus comes with many new capabilities for doing science (e.g. machine learning) and connections to the Big Data Market but at the same time with an enlargement of the HPC ecosystem. It is acknowledged within the EXDCI community that this new challenge will shape the future of Industry and Science. In this section, we first expose how ecosystem development has been addressed in EXDCI. Then we provide a short overview of EXDCI recommendations for the HPC Ecosystem.

The EU HPC Ecosystem is rich but complex. It gathers hundreds of organisms, research labs, universities, SME and larger companies; these stakeholders take part in numerous European initiatives and projects. EXDCI maintains a map of the entrire related European HPC Ecosystem (www.etp4hpc.eu/map-of-european-hpc-eco-system.html, for details on EXDCI, see Section 3.1.5). EXDCI addresses the Ecosystem coordination using two approaches. The first method is hierarchical via PRACE[42] and the ETP4HPC while the second is transversal.

The hierarchical method aims at producing roadmaps and at establishing the scientific challenges to be addressed. In EXDCI, this is the Strategic Research Agenda (this document) and PRACE Scientific Case [PRACE] respectively led by ETP4HPC and PRACE.

The transversal approach in EXDCI has been implemented by a set of events presented in the next paragraphs.

The first major initiative setup by EXDCI was the European HPC Summit Week that was held in Prague (2016) and Barcelona (2017). These events gathered hundreds of academics, FET-HPC[43], CoE[44], and company representatives.

The second major action has been to represent the EU partner in the international BDEC[45] initiative. The BDEC workshop series is an international initiative that allows scientists to exchange on road-mapping efforts in order to understand the paradigm shift underlying extreme data and computing. Previous efforts that started with the IESP[46] initiative have been very successful in helping the exascale roadmap definition. EXDCI has participated in the organisation of a set of BoFs and workshop at major events such as Supercomputing (USA) and ISC[48] (Germany). EXDCI technical workshops have been an opportunity to establish a strong link with the Big Data Value Association (BDVA[49]).

EXDCI's activities related to start-ups and SMEs focused on how the start-ups and the SMEs in the HPC area perceive their situation and what are the hurdles they face. We broaden the scope of the study with the point of view of the other stakeholders. This led to a set of recommendation promoting the partnerships between SME, start-ups, large companies and HPC centres.

Other transversal actions include monitoring the ecosystem by establishing a Balanced Scorecard to capture its evolution, synchronising with Eurolab4HPC[50], promoting HPC by publishing Career Case Studies. The latter aims at promoting the various interesting and exciting career opportunities that HPC can offer to young people.

The **non-technical** recommendations based on this approach are presented in Section 9.1 (Ecosystem Level Recommendations).

[41] European Exascale Software Initiative, **http://www.eesi-project.eu/**

[42] Partnership for Advanced Computing in Europe, **www.prace-ri.eu**

[43] Future and Emerging Technologies – High-Performance Computing

[44] Centres and Excellence in Computing Applications

[45] Big Data and Extreme-Scale Computing, **www.exascale.org**

[46] International Exascale Software Project

[47] Birds-of-a-Feather – an session organised at a Supercomputing Conference (SC, **www.supercomputing.org/**) aimed at sharing information on a certain area or aspect of Supercomputing

[48] International Supercomputing Conference, **http://isc-hpc.com/**

[49] **www.bdva.eu**

[50] **www.eurolab4hpc.eu**

3.3
THE INTERNATIONAL CONTEXT OF EUROPEAN HPC TECHNOLOGY

Outside of Europe, the international HPC scene is dominated by three ecosystems: China, US and Japan. Other countries have recognised the importance of HPC and have initiatives in place to develop their ecosystems. There is a clear trend to create synergies between HPC programmes and Big Data and Intelligence Artificial initiatives.

3.3.1 China

China has recently made a lot of progress in the number of installed HPC systems, in HPC technology and in HPC applications. The progression of China in the Top500 during the last years has been impressive. In 2016, China owns more than a third of these Top500 systems while two years ago only 7% of the Top500 systems were Chinese. The top two systems are installed in China and their combined performance (more than 125 Pflops) outperforms those of the other eight systems of the Top10 (less than 100 Pflops). In the field of HPC technology, the worldwide top HPC system is based on a processor developed in China. As a result, China now masters HPC processor and interconnect technology. With respect to applications, even if the Sunway TaihuLight system memory subsystem is not balanced compared to its processing power, a Chinese team has won the 2016 Gordon Bell award achieving impressive simulations with this system.

In its 2016-2020 plan, China has set the objective to achieve exascale. Currently, three competing projects are underway (Sunway successor, Tianhe successor and an industrial project led by Sugon). One of these will be selected to focus the effort to deliver an exascale system by the end of this five years plan. In this initiative, the convergence of extreme data and AI with HPC is one of the objectives and the architectures targeted should address the challenges of this new set of applications.

3.3.2 United States

HPC has a high political visibility and the HPC-related objectives are set by President Obama's Executive Order establishing the National Strategic Computing Initiative (NSCI) issued in July 2015. Nevertheless, due to China's development, the US leadership has been reduced. There are expectations to regain the lead in HPC systems with the installation of three top systems issued from the Coral procurement. Two systems (Summit in Oak Ridge National Lab and Sierra in Lawrence Livermore National Lab, both based on the OpenPOWER architecture) are planned for installation in 2017 and production in 2018 and will deliver between 125 and 200 Pflops each. The third system (Aurora), based on Intel KNH architecture, originally planned to be installed in 2018 in Argonne National Lab has been postponed till the delivery of an Exascale system in 2021.

In the future, the main efforts are organised under the Exascale Computing Project[51] (ECP) managed by the Department of Energy (DoE). ECP covers the development of technologies both hardware and software, systems and applications.

The first set of research contracts has been awarded by ECP for hardware technologies (six projects), software technologies (thirty-five projects), Co-design Centers fifty-five projects) and applications (twenty-three projects). Recently, the ECP has decided to add the objective of delivering an exascale system in 2021 based on an advanced architecture (capable exascale system are also planned for 2022-2023). The current roadmap is the following: DoE is used to award contracts for the acquisition of HPC systems that include a development phase and are signed three years before the delivery of the systems. The next systems will have to be designed to solve emerging data science and machine learning problems in addition to the traditional modelling and simulations applications.

[51] https://exascaleproject.org/

3.3.3 Japan

In Japan, the most important effort is the post-K project. This project plans to deliver an exascale class supercomputer in 2021. The project is managed by Riken[52] and includes the development of a new system architecture by Fujitsu, the delivery of a complete software stack and some advanced works in nine application domains. The approach adopted in the implementation of the post-K (called in the past Flagship 2020) is the following:

In relation to the system, Fujitsu develops a new processor-based on the ARMv8 instruction set architecture. This processor will implement the recently released Scalable Vector Extension to increase the performance of the system.

Besides the post-K project, Japan prepares new systems that will be used both for traditional HPC applications and Big Data applications. Japan has launched an ambitious plan in AI[53] (more than $1B) and the HPC community works on how to leverage HPC technologies for this field.

3.3.4 Other important ecosystems

Australia has an active HPC ecosystem organised around the National Computational Infrastructure. The research is mainly focused on application development and targeted a large set of scientific fields.

Brazil has an important program to develop HPC and to sustain both the academic and industrial development through this investment.

India has launched the National Supercomputing Mission to connect national academic and R&D institutions with a grid of over seventy high-performance computing facilities with a budget of Rs 4,500 crore (644M€). Some of the systems will be acquired from external vendors and some will be designed indigenously and used to develop India's own HPC technology.

Korea has significant HPC expertise and has decided to independently develop supercomputer technology. This effort does not target extreme performance but rather energy efficiency and artificial intelligence needs. Toward this end, the government will be providing 10 billion won (US$86.20 million) in funding each year for the next ten years.

Russia has developed a significant capacity to develop supercomputers and has installed petaflops systems. Nevertheless, recently Russia's progress seems slower than the others and its relative position is weaker.

Saudi Arabia has invested significantly in HPC and has now excellent skills located in top academic institutions such as KAUST.

South Africa effort in HPC is led by CHPC[54] that has installed a Pflops system in 2016. This ecosystem is focused on the use of HPC.

3.3.5 The International Collaboration of European HPC Technology and Application expertise

As a result of EXDCI Task 6.2[55] to date, there is an established and recognised presence of all the European HPC technology projects at the world's largest HPC-related conference (Supercomputing Conference[56], SC) in the form of a European Birds-of-a-Feather[57] (BOF) session organised annually, which attracts representatives of the international HPC arena. This task federates the efforts of European HPC in this area by ensuring a single interface for all the projects.

ETP4HPC (with the support of EXDCI) have produced a European HPC Technology Handbook[58], which includes up-to-date information on all the Projects – this document and other related continuously updated material is available on a web page dedicated to this task: www.etp4hpc.eu/euexascale.

ETP4HPC have approached the most prominent regions in HPC technology development in order to obtain updates on the work taking place in those countries. These actions will help the Projects develop their international collaborations as they mature and produce tangible results. International[59] partners can also access the ETP4HPC networking tool at http://www.etp4hpc.eu/networking.html in order to contact the members of the association.

[52] http://www.riken.jp

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

[54] https://exdci.eu/sites/default/files/public/files/4e-exdci-tech-workshop-2016-wp6.pdf

[55] http://www.supercomp.org/

[56] http://www.etp4hpc.eu/euexascale

[57] http://www.etp4hpc.eu/en/european-hpc-handbook.html

[58] The term 'international' in this document is used to refer to 'non-European' (i.e. foreign or overseas) projects from outside of Europe.

The key conclusions of our assessment of the international collaboration opportunities for the European HPC projects are as follows:

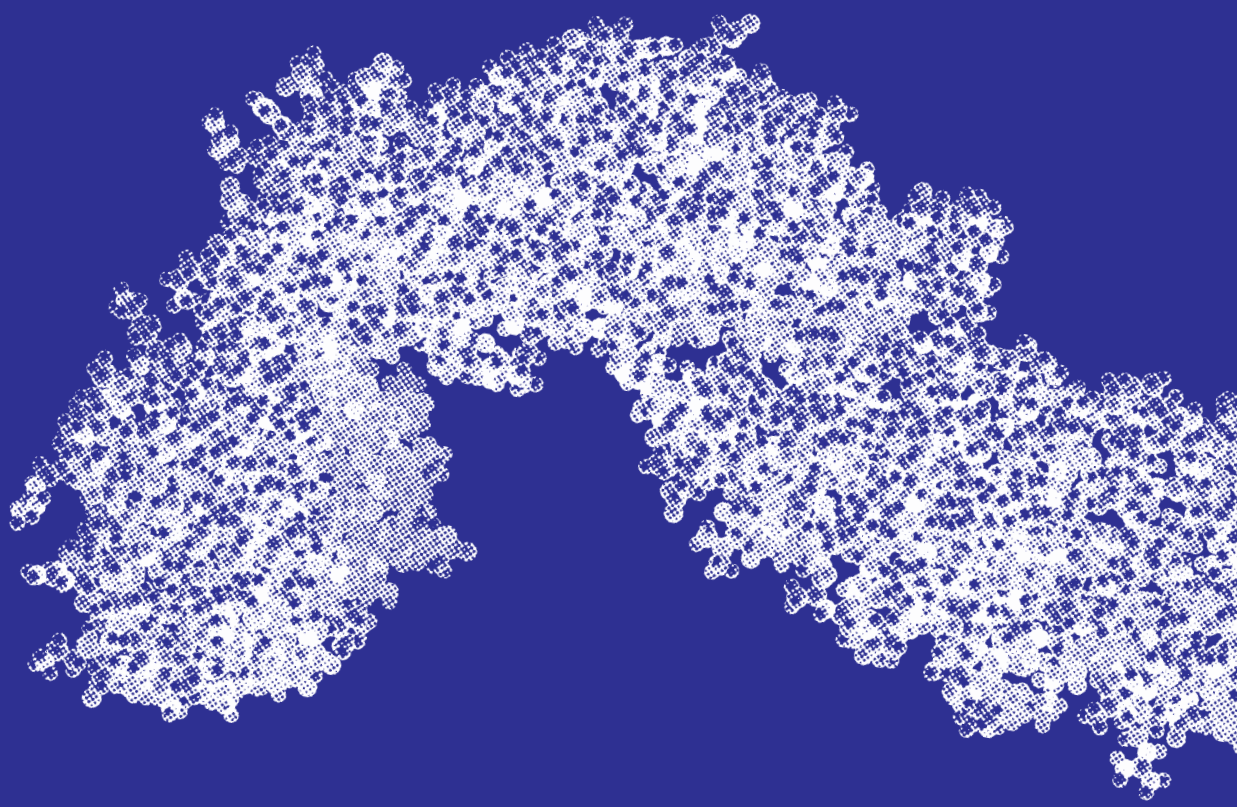
- The Projects are open to international collaboration opportunities and willing to engage in activities in order to facilitate this process. Likewise, the international community has demonstrated high level of interest in the results of the European projects.

- There are a number of areas where cooperation seems possible, and the Projects are able to identify these areas and pinpoint potential partners in both academia and industry. Some projects have already started work involving international partners.

- The European HPC Ecosystem should further facilitate this process by identifying areas of priority where European and overseas projects could jointly contribute to the goals of the international HPC community and organise e.g. common workshops in selected areas and research visits (in particular in the area of Programming Tools), leading to joint calls and other funding mechanisms. Also, a clear dissemination plan is needed in order to help the Projects reach the appropriate partners.


Silenne
VALSIN
move

4.

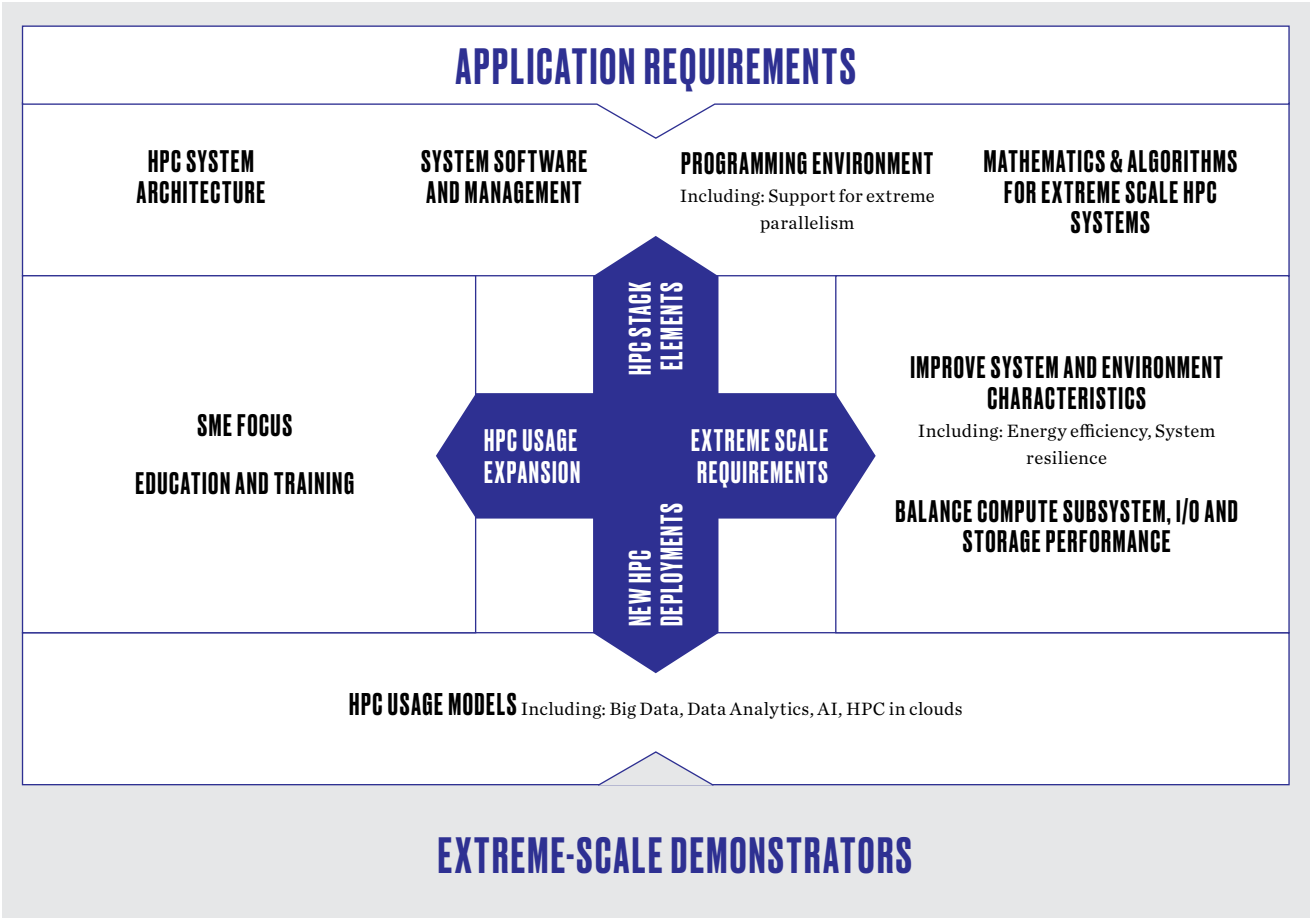


DIMENSIONS AND DRIVERS OF HPC RESEARCH

4.1
THE ETP4HPC SRA MODEL

The conclusions of the analyses above have been mapped onto our technological model developed in the course of the previous SRA issues. This 'Multi-dimensional HPC Vision' defined in SRA 1 (2013) remains the foundation of the SRA structure as a proven tool in dealing with multiple facets of HPC technology.

Figure 8.
The new modified four-dimensional model of European HPC technology development stemming from its first version as defined in SRA 1 (2013).



There is a demand for R&D and innovation in both extreme performance systems and mid-range HPC systems. Almost all scientific domains and also some industrial users require extreme-scale performance systems as soon as possible. There is also a need expressed in particular by industrial users and ISVs for more flexible, easier-to-use, more productive and cost-effective HPC systems delivering mid-range performance.

The ETP4HPC HPC technology providers share the view that in order to build a sustainable ecosystem, their R&D investments should not only target the top-of-the-range exascale objective. This market will be too narrow to yield a sufficient return on investment and support sustainable technology development. Such a strategy would weaken the European players. On the contrary, an approach that aims at developing technologies capable of serving both the extreme-scale requirements and mid-market needs can be successful in strengthening Europe's position.

As a consequence, the 'north' dimensions of this model points at the major areas of R&D in new technologies able to offer more competitive and innovative HPC systems for a broad HPC market, and the 'east' side at enhancing these technologies with the right characteristics to address the extreme-scale requirements.

The third element – the 'south' direction - is the current trend of developing new HPC applications. Besides traditional HPC workloads, an increasing amount Big Data/Artificial Intelligence/Machine Learning applications will need to be handled by HPC technology solutions. Also, the 'Internet of Things (IoT)' will change completely the landscape of using HPC technology, distributing the intelligence and control between the 'edge' and central systems. There is also a clear demand from some domains to use HPC systems for the control of complex systems such as smart grids. The Cloud delivery model is yet another trend affecting the features of future HPC solutions. Accordingly, the SRA has a dimension to address all these new usages (bottom – HPC Usage Models).

The 'west' dimension of the model refers to two factors affecting the growth and competitiveness of the European HPC ecosystem. A major concern expressed by the European HPC technology providers and the European HPC user community is the lack of skills. The recent years have proven that thorough, world-class expertise in any of the HPC stack levels and applications is extremely difficult to develop. Almost all HPC stakeholders suffer from the lack of skilled labour and the improvement measures undertaken do not meet the expectations. The technology provider Small- and Medium Sized companies (SMEs) require special attention as they face various hurdles in becoming acknowledged players in the HPC ecosystem.

SRA-3 brings further details on the ETP4HPC initiative called "Extreme-Scale Demonstrators" (EsD) and the model above reflects their importance. This approach described in Chapter 8 intends to proof and showcase the value generated by the H2020 HPC research projects in form of tangible results. These Demonstrators are a mechanism to integrate the work of the entire European HPC technology into working system prototypes, implementing elements from all areas shown in Figure 9 shown above.

Looking forward into the next years, the focus of the proposed research topics outlined in the following chapters will shift from a very technical computing centric view to a broader "extreme compute and extreme data" mix. This raises the question of cross-pollination between HPC and BD:

Cross-Pollination between respective BD and HPC platforms can support scenarios that require tighter coupling of compute-intensive analytics (BD) and data-driven simulations (HPC). Tighter coupling between HPC and Big Data can handle new and very challenging use cases emerging from the massive 'IoT-isation' of almost everything[60]. Indeed, real-time and complex interactions of billions of various smart things will require complex modelling and intensive simulations, which will be continuously refined through analytics and data-driven tuning of those models. This is turn generates continuous and dynamic changes in a complex symbiosis between the real, digital and virtual worlds.

A very promising example of this is the usage of Digital Twin[61] concept - used for improving the design and real-time operation of complex products/systems, e.g. for continuous monitoring and real-time optimisation of connected and autonomous cars. Indeed, by providing terabytes (TBs) of data per hour of operation, this complex system of systems scenario, operating in extremely dynamic conditions[62], requires (completely) new methods and tools for real-time understanding of data (e.g. situational awareness) and proactive (and extremely efficient) reaction in the case of anomalies.

[60] There will be 34 billion devices connected to the internet by 2020, up from 10 billion in 2015. IoT devices will account for 24 billion, while traditional computing devices (e.g. smartphones, tablets, smartwatches, etc.) will comprise 10 billion (http://www.businessinsider.com/jawbone-bet-big-on-fitness-trackers-and-lost-2017-7)

[61] Digital twin refers to computerised companions of physical assets that can be used for various purposes. A digital twin uses data from sensors installed on physical objects to represent their near real-time status, working condition or position. Digital Twins for product design have been in use for some time for highly capital intensive products such as jet engines and heavy machinery. However, as Digital Twins get more ubiquitous, democratised, and accessible, their benefits can be leveraged by every product manufacturer and product user (for self-service).

[62] 380 million connected cars will be on the road by 2021

We envision that the relationship for BD-HPC coupling to be dynamic and symbiotic. For example, by more speedily projecting the inferences from (big data based) real-time massive data streams into (HPC based) models and simulations, the temporal delta between as-designed and as-operated can be reduced considerably. This can impact product design, manufacturing, usage and servicing – opening up new business models and opportunities as we move from capital-expense business models to consumption-driven business models.

How can Big Data benefit from HPC?

Big Data applications are expected to move towards more compute-intensive algorithms for descriptive (data aggregation and mining) and predictive (statistics and forecasting) analysis. Prescriptive (decision making algorithms) analytics could be integrated with them to provide a feedback loop across the full decision-making process. HPC capabilities are expected to be of assistance for faster decision making.

How can HPC benefit from Big Data?

Analytics is expected to become a fully-fledged software component of the HPC ecosystem to process the massive results of large numerical simulations or to feed numerical models with complex data produced by other scientific tools. Iterative refinements of the models used by the HPC simulations could thus be done by benefitting from advanced data analytics tools and machine learning techniques. HPC can benefit from Big Data Management approaches, especially in the case of dynamic scenarios (HPC usually has the data close to processing, Big Data is much more flexible with the notions of data at rest, data on move, data in change).

The following part of the document deals with the dimensions set above. Section 5 deals with the latest technological trends, Section 6 describes the research priorities in detail and chapter 7 lists their milestones. Section 8 tackles the area of the Extreme-Scale Demonstrators. Section 9 includes an analysis of the non-technical areas.

4.2 AN EMERGING DEPLOYMENT CONTEXT FOR HPC

While in the forthcoming years technology for HPC systems will be further developed with a focus on extreme scalability, superb energy efficiency and resilience in the pursuit of Exascale use scenarios in the domain of technical computing, the fast growing Big Data market will demand the adoption and use of HPC technology to cope with high end HPDA and DL use cases. This will have a significant impact on the R&D priorities of HPC system providers.

As mentioned above, Big Data has gained a lot of momentum in the recent years, even to an extent that various groups of experts recommend to converge classical HPC and High-Performance Data Analytics software stacks. While each of these domains has its set of unique requirements in terms of the underlying IT infrastructure, there is an increased pressure for using more software stack elements in both domains and leveraging technology, methods and tools from across these domains.

On an international level, the BDEC initiative deals develops various approaches of 'stack convergence', as explained in the following chapter.

4.2.1 Paradigm splits in the context of data-compute

Two major splits have emerged in the today's global "big data"-"extreme scale compute" environments of today.

The **first split** concerns the divergence and the subsequent parallel evolution of two software stacks: one for big data, the other for high performance computing. This split was brought to the attention of the BDEC community early in the process[63].

[63] Reed, D. A. & Dongarra, J. Exascale Computing and Big Data Commun. ACM, ACM, 2015, 58, 56-68.

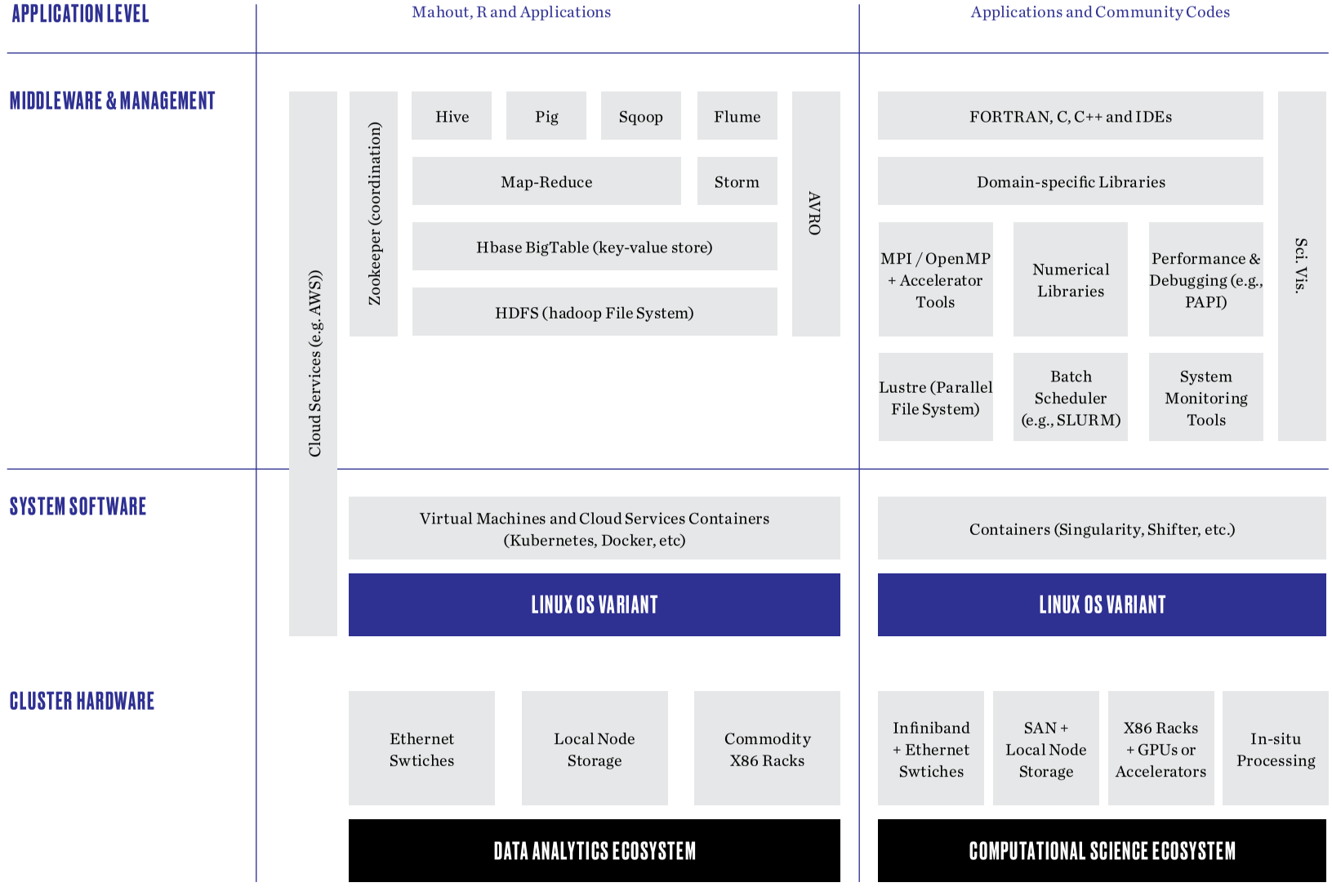


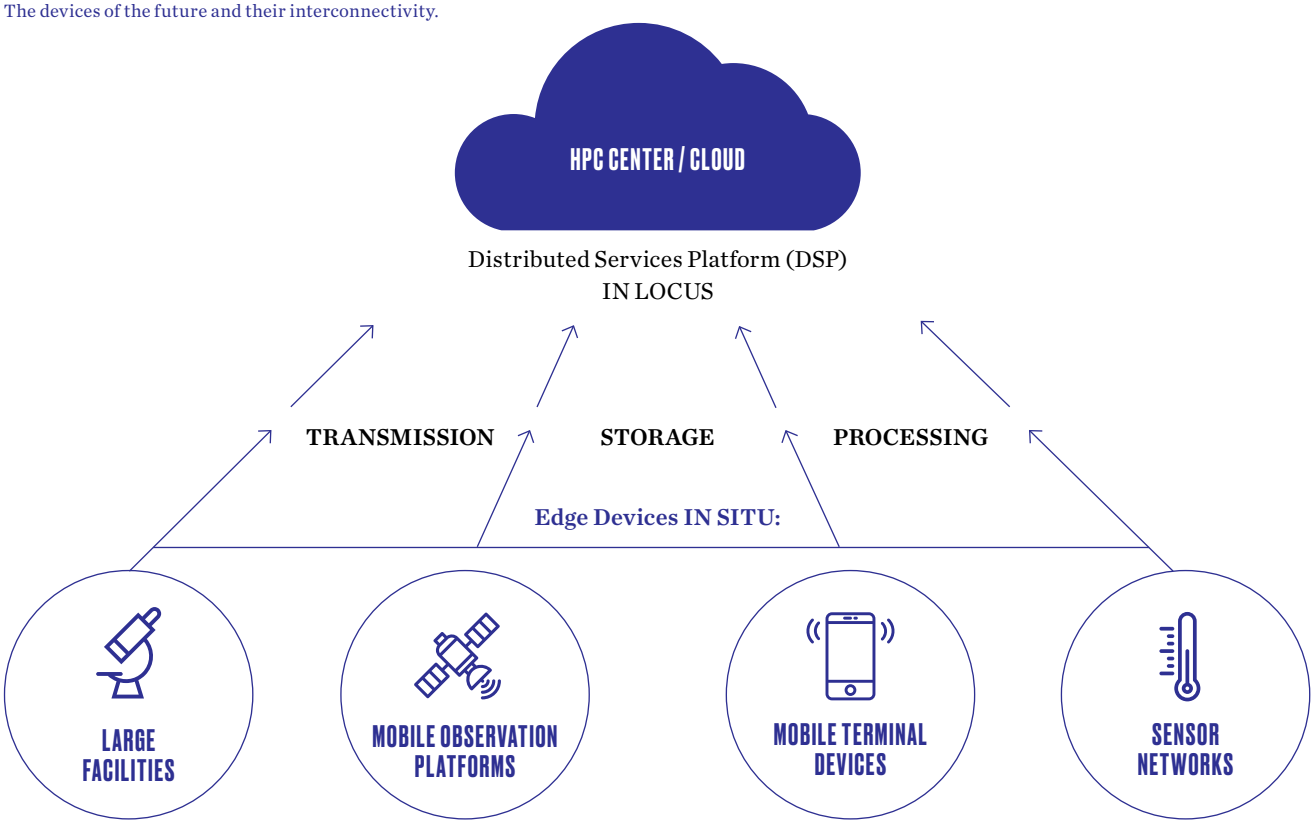
Figure 9.
The Data Analytics and Computing stacks side-by-side.

The *Computational Science Ecosystem* (on the right) developed and flourished over the course of roughly four decades (primarily) to increase the capabilities of scientists to model and simulate, i.e. to enable scientists and engineers to project, in more detail, the consequences of theories that had been or could be expressed mathematically. Meanwhile, the rapid growth of the Data Analytics Ecosystem (on the left) has occurred largely during the last fifteen years. For the most part, however, it is not being developed by the scientific computing community to explore the rising flood of data from new instruments and sensor systems, but rather by an equally thriving group of academic and commercial software developers to process the torrents of business, industrial process, and social network data now being generated by consumer devices and the burgeoning Internet of Things. The pace of change in the data analytics ecosystem is extraordinary, already rendering obsolete some of the elements in the figure above. In fact, there is probably a secondary split occurring, into a new "Machine Learning" stack.

The second split, more fundamental, has to do with the need for a new "internet" that will reconcile and bring together the distributed internet of things, out on the "edge", with the actual centralised model of cloud and high performance computing.

Even the ultimate convergence of the HPC and High Performance Data Analytics (HPDA) ecosystems, could it be achieved, would not help with the ongoing breakdown of this other, more basic paradigm, namely, the one in which networks only forward datagrams while all other storage and computation is performed outside the network. The problem is that much if not most of the explosive growth in data generation today is taking place in "edge environments" i.e. *across the network and outside both HPC and Cloud data centres.* This includes not only major scientific instruments, experimental facilities, and remote sensors (e.g., satellite imagery), but even more importantly, the generators of an incredible welter of digital data coming from the "Smart Cities" and "Internet of Things" (IoT) concepts. Thus, this remarkable reversal of the direction of the data tide, which turns the familiar "last mile problem" into a multidimensional "first mile problem", represents a challenge for which neither cloud-based HPDA, nor centre-based HPC have a solution. In fact, explosive growth in data generation in edge environments seems to clearly indicate that revolutionary innovation in distributed computing systems is becoming an increasingly urgent requirement. We believe this represents the breakdown of the bipartite cyberinfrastructure paradigm that has been dominate dominant for nearly three decades, making the problem of 'convergence' substantially more complex and momentous.

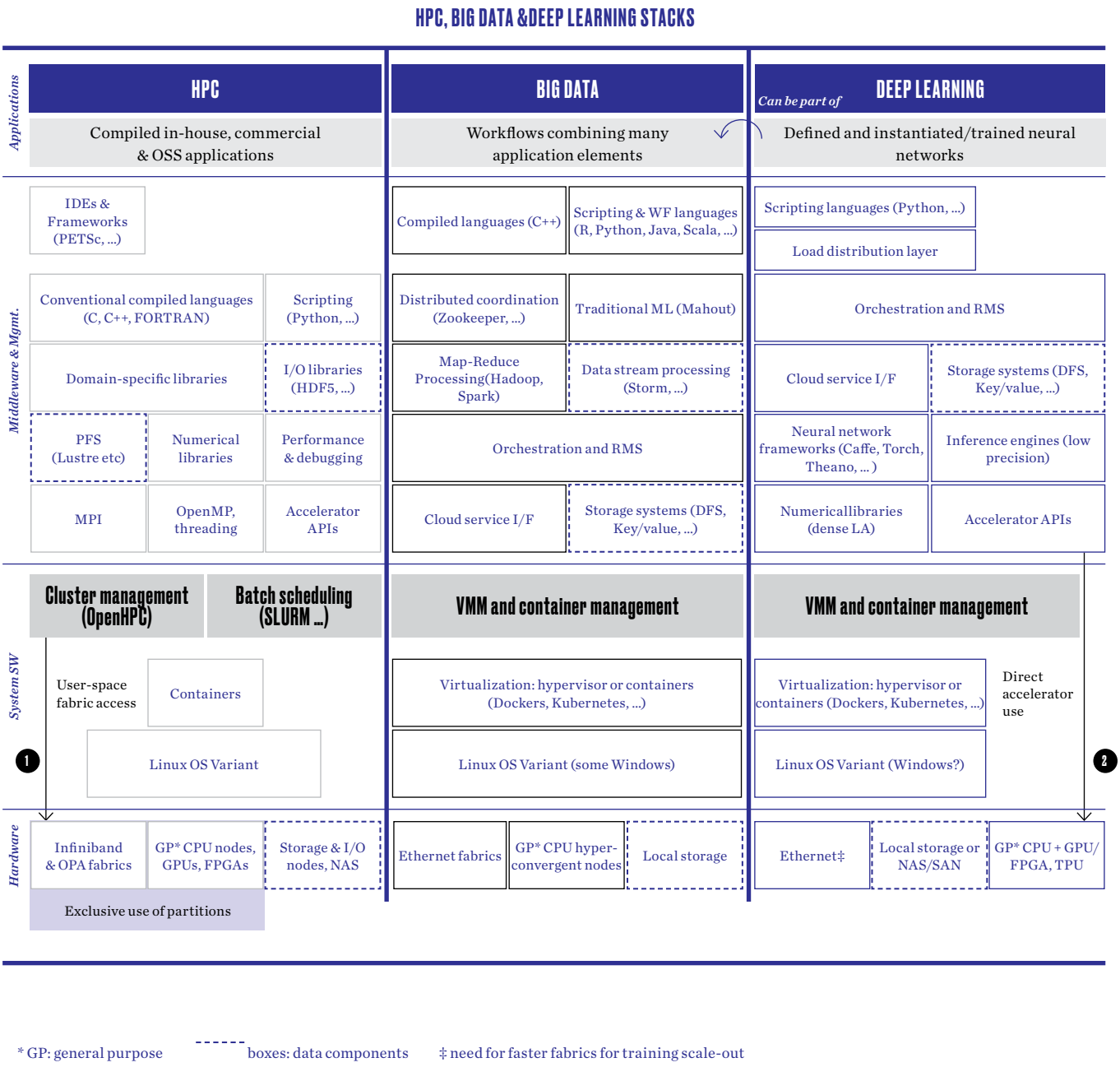
Figure 10.
The devices of the future and their interconnectivity.



4.2.1
Extending and updating the stack models

The view expressed in the previous section arose from BDEC-related discussions and EXDCI workshops on the theme that took place over the last few years. The more recent developments towards more complex scenarios led us to re-approach the software stack issue and produce an extended view. This is detailed below, with a summary of the salient features listed per software stack (column).

Figure 11.
The three stacks side by side. The paragraph below explains the contents of the three columns.



* GP: general purpose ------ boxes: data components ‡ need for faster fabrics for training scale-out

Supercomputing Column:

- Use of scalable and high performance (latency plus BW) interconnect fabrics for performance reasons
- Use of dedicated compute (often with accelerators attached) vs. storage nodes vs. "login" nodes
- Each single workload run on multiple/many nodes (up to 10000s) in a tightly coupled fashion (usually in a spatial partition of the larger system)
- Linux is the only OS, containers are being taken up, VM/Hypervisor style virtualisation is out
- Direct use of fabric in user space is required to keep latencies down
- Batch scheduling works on 100s of large, parallel jobs (10s through 10000s nodes each); this is a large difference to what "scheduling" means in Big Data & Deep Learning
- Data kept in parallel file systems and accessed using Posix semantics
- I/O libraries are used to encapsulate data formats and organisation
- Conventional, compiled language rule; scripting languages like Python are making inroads
- Limited use of higher level frameworks and IDEs – canonical example is PETSc for PDEs and ODEs
- Applications are compiled into binaries and distributed/executed in the "old-fashioned" way
- Use of HPC applications via Cloud interfaces is emerging

Big Data Column[64]:

- Ethernet use prevalent (economic and 'good enough')
- Nodes are "hyperconvergent-" – i.e. one kind of nodes fulfils all functions (compute, data storage, access)
- Virtualisation (VMs or containers) and VMM are prevalent
- Multitude of storage systems: PFS/DFS, key/value stores, object stores
- Orchestration and resource management/scheduling is fine-grained (at a level of nodes, not large partitions)
- Two prevalent application patterns: Map/Reduce and Data Streaming, with frameworks for each
- Non-Deep-Learning ML plays a significant part
- Since applications are usually complex workflows, distributed coordination is important
- C++ plays a role, but scripting or interpreted languages are very strong

Deep Learning Column[65]:

- Deep learning training uses large nodes (most often with several accelerators, GPGPUs, FPGAs and special accelerators like TPUs)
- Fabric is Ethernet, with current R&D looking scale-out over high-performance fabrics
- OS & virtualisation are the same as for Big Data, however direct access of accelerators is prevalent
- Numerical libraries are key – DGEMM/SGEMM, gradient descent and "tensor" data manipulation determine performance; additional numerical algorithms like FFTs will likely become important in the future
- For training & inferencing, a multitude of high-level frameworks are in use, with new ones bubbling up from time to time; training is highly compute (and communication) intensive
- Inferencing uses much less compute & communication, and it is regularly done at reduced precision (FP16, 8-bit, sometimes even less)
- Data stores are the same as for Big Data
- Applications use mainly scripting/interpreted languages, with some C++
- Deep learning is usually a step in a large Cloud or Big Data workflow

[64/65] Big Data and Deep Learning workloads are mostly running in cloud environments.

4.3
APPLICATION REQUIREMENTS

What are the drivers for the technical priorities in HPC research and which application and usage-oriented aspects are the main influencers? The purpose of HPC technology is to serve HPC applications (which in turn serve the needs of the scientific, industrial or societal user as described above). This section lists the main high-level requirement trends key areas of use impose on underpinning HPC technology.

4.3.1
Scientific Application Challenges

Europe has a particularly strong role in scientific and industrial applications with a large fraction of the applications used in the world being originally developed in Europe. In their transition towards Exascale, applications face a number of challenges and in this section we highlight some of the key challenges based on input from the EXDCI project[66] and the current Centres of Excellence for High Performance Computing (CoE)[67]. It is important to note that this list is a selection and by no means comprehensive, neither in terms of the domains potentially using Exascale systems, nor the challenges identified.

While different areas have specific challenges we highlight below, there are a number of cross-cutting issues that affect all areas:

Portability and maintainability are a key concern for applications, also when approaching the Exascale. It is therefore mandatory to have well defined, stable, and portable interfaces, where possible relying on standards, particularly when it comes to new developments such as heterogeneous architectures, deeper memory hierarchies, and new networking concepts. Furthermore, there is a need for improved software engineering practices including testing for both correctness and performance, collaborative software development, benchmarking and validation, and code re-factoring and modernisation.

The increased complexity of future HPC architectures is making productive coding and performance tuning an even more difficult task. Novel approaches e.g. Domain Specific Languages (DSL) relying on smart underlying system software layers can help abstracting the complexity of future HPC architectures.

Scaling towards the Exascale will increase the data requirements drastically. Application will need high bandwidth memory and networks, as well as new techniques to deal with data in a more energy and performance efficient way, such as in-situ and in-transit post processing coupled with smart AI techniques based on machine/deep learning for automatic features detection/tracking for example. In that sense, efficient cohabitation between HPC, big data, and AI software will also be needed.

Knowing the hardware and software constraints brought by Exascale architectures, the need to optimise both performance and energy consumption and the vast number of applications to consider, it appears important to develop low overhead, accurate (high resolution) and possibly automatic sampling tools, able to profile performance, power consumption and data movement in a scalable way. Impact on application performance should be minimised (less than 2% overhead). Information collected could feed a central database of thermal and performance signatures allowing the user to know exactly the behaviour of their application.

The increased size of Exascale systems will likely increase their exposure to faults. Fault tolerance and resilience techniques, both on the algorithmic/solver level as well as on the runtime/programming environment level will be required.

Finally, as Exascale will allow not only to support capability based applications able to scale out on a full machine, but also capacity based applications based on coupled multi-physics and multi-scale approaches, it is mandatory at the European level to develop unified scalable environments allowing to support complex end-to-end workflows, smart coupling of applications, the assessment of uncertainties quantification (UQ) or perform large optimisation studies.

While most Exascale efforts happen in the context of open source software developments, many ISV codes suffer from a lack of scalability and difficulties to embrace advances made by open source developments. This is a particular challenge for certain domains and particularly industry relying on ISV codes. Specific efforts are needed to improve ISV codes, including training, profiling, refactoring or co design.

[66] European eXtreme Data and Computing Initiative: http://www.exdci.eu

[67] In 2015, a first set of nine CoEs have been created to provide efficient and scaling software and support scientific communities in their use of the software and the European HPC infrastructure. These CoEs cover currently the areas of Materials, Energy, Biomolecular research, Weather and Climate, Biomedicine, and Global Systems Science. A transversal CoE deals with performance optimisation and productivity. Through the work programme 18/19 the CoE concept will be further supported and expanded to areas not yet covered, such as Engineering and Big Data.

	REQUIREMENT	CHALLENGE
A1	Portability/Maintainability	Development of standards for programming heterogeneous architectures, deeper memory hierarchies, and new networking concepts; DSLs with smart system software layers
A2	Automatic profiling of applications	Develop automatic lightweight and accurate profiling tools for taking into account performance, power consumption and data movement of applications
A3	Analysing large amounts of data	In-situ and in-transit scalable analytics using e.g. ML/DL techniques
A4	Cohabitation of different software stacks	Allow classical HPC software to interact with Data Analytics software stacks
A5	"Separation of concerns": refactoring applications to separate the parts that are specialised for different HW.	Encapsulate kernel functionalities in self-contained HW-specific modules that should become part of domain specific libraries, or kernel system libraries when common to other application fields. The latter should be optimised, tuned and co-developed across different domains, and co-designed with HW vendors and specialists in performance optimisation and exascale programming paradigms.

4.3.1.1 Industrial and engineering applications:

The use of advanced numerical simulation and HPC, traditionally reserved to few industrial domains such as automotive and aeronautics is now encompassing a wide range of domains including nuclear energy, oil & gas, combustion, renewable energies, chemistry, to name a few, from large companies to innovative SMEs. It is a strong asset of Europe's in terms of the companies already involved or potentially involved and a strong reservoir of growing competitiveness for Europe.

In Aeronautics, in order to meet the challenges of future aircraft transportation ('Greening the Aircraft'), it is indispensable to be able to flight-test a virtual aircraft with all its multi-disciplinary interactions in a computer environment and to compile all of the data required for development and certification with guaranteed accuracy in a reduced time frame. Such challenges include HPC and data requirements in terms of:

- Increasing the scalability of individual CFD, structure, combustion, acoustics- simulation codes by working on new scalable numerical solvers
- Developing automatic grid generation tools for handling complex geometries
- Next generation of couplers for multi scale and multi physics applications (aerodynamics, structural mechanics, aero-elastics, flight mechanics, aero-acoustics, engines)
- Next generation of uncertainties/optimisation framework for multidisciplinary aircraft design based on high-fidelity methods
- Handling and visualisation of Big Data with the development of in-situ data analysis

In the automotive industry, the major use of HPC is dedicated for large-scale CFD studies (aerodynamics, combustion) and massive crash test optimisation studies. Some companies state that today a crash study over one year is generating around 40 exabyte of rough data which is reduced on-the-fly to a volume of 400 TB where data analytics is performed. They are expecting a big increase of the size of the rough data generated, due to use of composite materials and enforced safety rules, and try

also to use the temporal data during the on-the-fly reduction phase to enrich its data analytics.

The issue of developing in-situ data analytics with tools for automatic structure (or pattern) detection will become key in the future for supporting engineers when diving into such a huge amount of data.

In Oil & Gas, HPC is mainly used for the development of efficient and accurate seismic processing methods for exploration and production (4D seismic coupled to reservoir modelling). Some Oil & Gas companies have a strong Exascale roadmap toward the use in 2020 of Full Wave Equation methods on massive amount of data acquired at high frequency (in the range of 3 to 55 Hz), requiring compute nodes based on heterogeneous/many-core resources, large amount of fast local unified memory (>32 GB and 300 GB/s HBM[68]) and high bandwidth interconnect (without PCIe[69] overheads) and I/O capacities. Moving from current TTI RTM methods (tilted transversely isotropic reverse time migration) to advanced L2ERTM (least squares/residual shot elastic RTM) will require a 500 to 1000x increase of performance regarding today x86 scalar systems.

In this context, software requirements in the field of ultra-scalable solvers, in-situ post-processing with compression of data, and tools for handling uncertainties quantification are crucial to develop.

Nuclear energy is using intensively CFD[70] methods (Large Eddy Simulations / LES / and quasi-Direct Numerical Simulation / DNS / methods) and Monte Carlo neutronic transport for the improvement of safety and efficiency of the facilities (especially nuclear plants), optimisation of maintenance operation and extended life span. Immediate needs in the field require (unstructured) complex meshes in the billion-cell range. Studies in the near future could easily an order of magnitude more. Such studies will require in the range of several hundred thousand (or millions) of cores during several weeks.

Finally, combustion is used to produce around 90% of the earth energy and is essential for ground and air transportation, electricity production, industry applications or safety. Efficient analytical methods are required to develop models from highly-resolved DNS data for cheaper but also predictive simulation types such as Large Eddy Simulations (LES). Subdomain simulations allow for tackling highly complex systems such as combustion chamber/multi-stage turbines with increased modelling accuracy using multi-physics code coupling for fluid/solid interaction for both structure dynamics and thermal response for example.

Fully resolved complex chemistry in real engines is still out of reach from current HPC platforms, however analytically reduced chemistry models allow today to predict global tendencies for emissions predictions with reasonable computational cost increase (x3 compared to simple models). Additionally, by its nature, combustion involves highly intermittent phenomena which entails require long time-resolved simulations. This translates to generating large amounts of data where data mining will become crucial.

	REQUIREMENT	CHALLENGE
IND1	Ultra scalable solvers	Communication avoiding methods, parallel in time, encapsulating space and time combination of high order schemes (like DG) with implicit or semi-implicit time stepping allowing large time steps.

4.3.1.2 Weather and Climate:

The general development in Earth-system modelling for both weather and climate science is towards finer scales and larger ensemble sizes to provide more fidelity in the predictions and in particular the representation of high-impact events. The community targets kilometre-scale simulations of the global atmosphere and oceans with ensembles of hundreds of members and a simulation speed of several years per day.

This trend imposes scalability and operability limits on weather and climate prediction centres that need to be addressed through fundamentally new scientific and technical methods, namely numerical techniques supporting domain-distributed computations, including parallel-in-time methods, and new programming models and domain-specific languages separating readable science code from highly optimised libraries with efficient, parallel code and specific implementations for several target computing architectures.

[68] High-Bandwidth Memory

[69] PCI Express (Peripheral Component Interconnect)

[70] Computational Fluid Dynamics

For computing, the key figure is power consumption per forecast and time to solution, while for I/O it is the absolute data volume to archive and the bandwidth available for transferring the data to the archive during production, and time-critical dissemination to multiple users. Today's output is of O(2)[71] Pbytes per week in weather forecasting, and this is expected to grow by at least one order of magnitude by 2022. Both aspects are subject to hard limits, i.e. capacity and cost of power, networks and storage, respectively.

The urgency of the adaptation to highly parallel computing is different for each component of the forecasting system, namely data assimilation, forecasting and data post-processing/archiving. Despite ambitious targets being set for model resolution, complexity and ensemble size, today the bulk of the calculations are not performed with configurations that utilise the maximum possible number of processors. Data assimilation, extended range prediction and research experimentation mostly operate at relatively low resolutions, predominantly for affordability reasons. However, the operational forecast suites also contain cutting-edge components that fully exploit current HPC capabilities.

	REQUIREMENT	CHALLENGE
WC1	Disruptive numerical methods	Communication avoiding methods, parallel in time, encapsulating space and time, combination of high order schemes (e.g. DG) with implicit or semi-implicit time stepping allowing large time steps. Optimised communication for advection schemes; mixed-precision methods & AI techniques
WC2	Efficient coupling of capacity and throughput computing	Perform ultra-large O(100-1000) ensemble calculations implying strong I/O systems with novel storage layouts (beyond POSIX) and support to resilient (fault tolerant) end to end workflows

[71] The order of magnitude = 100

4.3.1.3 Biomolecular Research:

Biomolecular research covers different areas, including quantum mechanics/molecular mechanics, molecular dynamics, and biomolecular modelling.

Biomolecular modelling: The human genome has revealed over 20,000 expressed proteins, the proteome, which are the workhorses of our life, performing all kind of critical functions in our cells. In humans, this interactome (interactions between each proteins) consists of hundreds of thousands of dynamical protein-protein (and other molecules) complexes. Miscommunication is this complex network can be at the origin of diseases, which is why it is important to understand how this network works at atomic level. This requires thus adding the 3D structural dimension to it. Considering the size and complexity of this network, it is clear that experimental methods only will not be able to provide all answers. This is where biomolecular modelling and in particular docking can play a crucial and complementary role. There are two main challenges requiring exascale computing:

1. Adding the structural dimension to the hundred thousands of interactions.

2. Given the knowledge of the proteome, predicting how the interactome will look like.

These require hundreds of millions of docking runs generating hundreds of exabytes of data.

Molecular dynamics is powerful tool that can provide insight in molecular processes in atomistic detail. Due to short time steps (femtoseconds) compared to the time scales of biologically relevant transitions, which are on the microsecond to second range, an enormous number of integration steps is required. The computational cost is dominated by the non-bonded pair interactions, which can be computed efficiently on modern SIMD or GPU hardware. Simulating biomolecules is a strong scaling problem, because the system size is fixed (and larger systems come with larger time scales). A simulation of a typically sized system of 200 000 atoms takes about 10 exaflops per microsecond of simulation time. A single simulation is not an exascale problem. But nearly all problems of interest involve calculating distributions or free energies and how these vary with e.g. different ligands and/or protein mutations, which requires more sampling and increases the cost by

orders of magnitude. In current large-scale studies the number of combinations times the simulation time needed per case reaches a zettaflops.

QM/MM Free energy calculations. A myriad of biologically relevant processes requires a quantum mechanical descriptions. First principles (QM) free energy calculations are excellent HPC applications: studies employing simulations with ~1.5 million processor cores for systems of about 2,000 atoms are available in recent literature. However, most realistic systems are much larger. Then, the method of choice is arguably the so-called hybrid QM/ MM. Typically, the QM and MM parts consist of ~102 and ~104 atoms, respectively. In most cases, the QM problem is treated at the density functional theory (DFT) level while for the remaining part force fields are used. This setup may take up to 20 petaflops per time step of dynamics and consequently it requires order of zettaflops in practical applications. Replica-based simulations, keeping linear scalability, are excellent QM/MM free-energy candidates for exascale machines. This will require that (i) hardware fault-tolerant algorithms be developed as outlined above; (ii) scaling DFT bottlenecks, such as 3D FFT and two-electron integrals, be addressed. Most importantly, improving load balancing algorithms for exascale applications will pave the way to the description of diffusion processes over the QM/MM interface, which are of paramount importance in biophysics. Indeed, these processes require advanced theoretical frameworks such as the Hamiltonian adaptive multiscale scheme, for which load imbalance turns out to be critical.

	REQUIREMENT	CHALLENGE
BM1	To deal with millions of independent jobs	Efficiently run a huge number of jobs deal with their input and output data
BM2	To efficiently simulate the long-range electrostatics treatment.	Efficient and scalable 3D FFT or alternatives
BM3	Simulating diffusion processes within a hybrid resolution (multiscale) approach.	Adaptive load balancing schemes

4.3.1.4 Energy

The goal of the energy CoE is to improve means of production, storage and distribution of clean electricity. This involves areas as diverse as meteorology, where very short term forecasting is needed to predict the production of solar and wind farm and their efficient coupling to the grid and energy trading; fusion for energy, where coupling kinetic and fluid codes is necessary to model the entire chain of processes from vessel core to edge; discovery and design of new energy materials for photovoltaic cells, batteries and super-capacitors; and energy hydrology to manage geothermal and hydro-power including the influence of climate change on these resources. Key applications identified as having a high potential to exploit exascale and which would benefit from reengineering efforts include Gysela (fusion), Parflow (Water), Alya (Meteo), and Metawalls/ BigDFT/PVNegf (Materials).

Exascale computing will enable significant step changes in the predictability and management of renewables as their share of the energy mix increases towards 100% over the coming decades. There are a number of specific challenges which arise in this domain for wind, solar, hydro and fusion power which we focus on here. For example, a single large eddy simulation of turbulent flow through 100 turbines of an entire GW-scale onshore wind farm with complex terrain geometries would require billions of grid points and millions of time steps; the whole thing then repeated for a series of meteorological conditions to obtain an overall power output estimate. Similarly, accurate hydropower prediction relies on the combination of physically-based terrestrial water-for-energy models with observations providing the current state of the hydrologic states and fluxes. Resolving the plasma turbulence that governs the performance of a nuclear fusion reactor from electron scale ($\sim 10^{-4}$m) up to ITER size (~1m) with realistic time steps ($\sim 10^{-7}$s) over an energy confinement time (~1s) requires exascale. At the grid management level, output from the physical models of intermittent power sources must be linked to observations via ensemble calculations, resulting in a probabilistic power forecast needed to ensure stability and predictability.

	REQUIREMENT	CHALLENGE
E1	Meteo or hydro modelling using combined elliptic/ parabolic equation systems (e.g. incompressible Navier-Stokes equations for wind turbines, gyrokinetic or full Maxwell equations for fusion application)	Scalable algebraic solvers (e.g. multigrid), capable of exploiting accelerator hardware (see also WC1)
E2	Efficient coupling of capacity and throughput computing	Perform ultra-large O(100-1000) ensemble calculations to generated probabilistic power forecasts; batch systems and resource managers tuned for this coupling

4.3.1.5 Fundamental Sciences

In astrophysical fluid dynamics (and its geophysical counterparts), high performance simulations are devoted to the understanding of multi-scales, multi physics systems such as the interstellar medium, convection and turbulence in stars and planets, dynamo action and magnetised (low plasma beta) dynamical systems, global instabilities, disk accretion. By 2022, it should be possible to simulate systems that are at least 4 orders of magnitude larger in each dimension on a regular basis in order to do a systematic parameter space exploration, with the most extreme grand challenge simulations reaching about 4.5 to 5 order of scale difference in each direction. The largest simulations in astrophysics and geophysics to model either geo or solar dynamo, interstellar medium structuring or galaxy mergers required huge resolution to resolve the global and small-scale dynamics. This translates into hundreds of terabytes of data per simulations. Huge data management of 4-D structures (space + time) are required to understand the complex nonlinear physics and feedback among the various scales/objects/processes. Immersive (remote) data visualisation is required to identify key structures. Overall large memory per core (4 GB/core or beyond) and high bandwidth must be accessible as the global nature of many of the astrophysical problems make it difficult to run on slow/low memory systems. In the field of laser-matter interaction the commissioning of 3 large-scale facilities under the European Light Infrastructure (ELI) project will likely drive a huge increase in demand for heroic Particle-in-Cell simulations. Currently 3D, multi-billion particle simulations are routine and are already capable of matching experimental conditions for some laser-electron schemes. However, ion acceleration schemes that rely on denser material are still woefully under-resolved, and numerical results are often too optimistic regarding the beam properties. To achieve quantitative predictive power in this case, at least 10-100x more particles would probably be necessary.

Finally, in the field of High Energy Physics and QCD, the goal by 2021 and beyond is to perform most lattice calculations of hadronic systems at or near the physical pion mass, with lattices representing physical volumes of (4 fm)3 and larger. To achieve robust signals from these types of calculations, the scale and of the problem must be increased by at least a 1000-fold compared to today's calculations, and most likely larger.

In terms of software, the available simulation codes for simulations in Lattice QCD (LQCD) are highly advanced. The community produced and maintains libraries comprising computational kernels that can be used by a variety of applications in this domain. These are highly optimised for a given processing architectures. Examples are QPhiX[72] and QUDA[73] that are optimised for Intel Xeon Phi processors and NVIDIA GPUs, respectively. Optimised communications libraries exist as well, which utilise the low-level interfaces to the hardware to cut down latencies and to optimally use the hardware capabilities. These software packages are maintained by a broad community that is and will be willing to invest into optimised software also in the future.

	REQUIREMENT	CHALLENGE
FS1	Immersive (remote) visualisation	Develop scalable post processing tools able to visualise remotely in an immersive way massive amount of data
FS2	Extreme level of scalability and optimal exploitation of the available compute resources	A system software that facilitates optimisation close to the hardware to allow, e.g., for very low-latency communication

[72] https://jeffersonlab.github.io/qphix/

[73] http://lattice.github.com/quda

4.3.1.6 Global System Science

The exascale challenge of Global System Science (GSS) is not scaling up a single application. This community is a rather new community, thus the applications per se are still on a small-scale stage. However, there is a need for this community to extend their applications and the input data sets to other disciplines. Refugee stream simulations, for example, require also knowledge about economics, climate and other disciplines. Thus, GSS applications can be part of a wider workflow of interactions of different simulations (a so-called multi-sim approach) of which each may generate big datasets. As this might affect the execution on a grid of HPC resources (e.g. as there is natural network bandwidth limitation), GSS applications provide the requirement of running most of the workflow parts (potentially iteratively) on one single system. Thus, this system needs (a) a high core count, (b) highly efficient and high-speed networks and (c) data management/data analytics resources on the hardware site but also adapted system software for resource management/scheduling and network optimisation.

	REQUIREMENT	CHALLENGE
GSS1	High-speed data transfer of diverse data sources, e.g. data streams but also of the results of simulations in a workflow execution of different simulations.	Develop scalable post processing tools able to visualise remotely in an immersive way massive amount of data
GSS2	Providing the necessary means to execute a workflow (of partially parallel running) simulations, with potential strong dependencies on output data of simulations within the workflow.	Develop complex data management techniques to couple HPC and HPDA components to efficiently process data sets (generated and from external sources such as streams)

4.3.1.7 Materials

The computational study of materials has reached a stage where *predictive* approaches based on a quantum mechanical description of electrons and nuclei can be brought to address systems of unprecedented complexity and new classes of properties and spectroscopies. By combining *high performance and high throughput* computing strategies, they are leading to a new paradigm in the design of materials and nanosystems which can accelerate discovery, science and innovation in all fields. Major impacts are expected in all fields of research and technology where inorganic, organic or biological matter and structures play a critical role, and in designing the evolution of manufacturing as a whole. A key challenge will be the predictive simulation and design of nanoscale devices that will characterise the evolution of information technology 'beyond-Moore'. In turn, the recognised strength of Europe in methods and codes in the materials domain can be an asset in HPC developments towards the exascale in terms of co-design and frontier use cases.

Materials simulations produce also large amounts of data that are valuable to be stored, shared, identified (via Digital Object Identifiers, DOIs), and analysed. The experience with big data analytics shows the necessity of domain specific approaches in materials science and engineering. The properties of materials are manifold and intricate. Pre-knowledge is crucial for applying data mining techniques such as compressed sensing, subgroup discovery, etc., contrary to a pure machine learning approach. Tailored databases, designed for efficient data mining of heterogeneous results and able to store the full provenance of each object as well as workflows, were developed and will be further expanded and widespread. They are especially instrumental in the context of high-throughput computing (HTC), which requires massive automation of complex sequences of simulations and the related automated interactive tools.

	REQUIREMENT	CHALLENGE
M1	Enabling high-throughput calculation, automatic storage of data, and sharing of data and workflows	Workflow and data management systems
M2	Improve basic or kernel libraries, e.g. linear algebra and FFT, to optimise for the specific sizes that are needed. This requires co-design and closer interaction with library developers.	Better parallel implementation of 3DFFT (or better HW supporting its memory/communication pattern), new iteration loop for dense Hermitian matrix eigenvectors.

4.3.2
Future trends in the industrial use of HPC

As of now, the use of HPC by industrial companies is one of the most important pillars of digital product design. HPC is making its way into decision-support processes and acts as the computational backbone of cyber-physical systems.

However, depending on the application-field, the current bottlenecks can vary:

- Industrial fields such as aerospace, electrical energy and oil and gas are essentially in the same public/academic research area: their dominant applications are based on either in-house or partner codes, for which the source-codes are available. Furthermore, these industries are aiming at simulations that definitively require more compute power than available today by orders of magnitude. These are ideal partners in the development of HPC industrial use at exascale.

- Other areas such as automotive, chemical, pharmaceutical and general manufacturing industries rely nearly entirely on ISV-codes; the corresponding end-user organisations do not have access to the source codes. Even if much higher compute-power is required, improvement and thus a path to exascale can only be established together with the ISVs. However, a significant part of the end-users might be less interested in speeding up individual simulations, but rather achieving higher throughput in parameter-studies, optimisation etc., which makes this process more complex.

Therefore, it is necessary to start or continue strong performance-engineering efforts with industries falling into the first category and with important ISVs serving the second category. Apart from the ′classical′ use of HPC for numerical simulation, the development in the industrial HPC markets is moving disruptively towards novel usage patterns which are not yet visible in public science and research. Hyperion [Hyperion] and other research studies confirm that one third of the newly installed HPC-power in industry is not used for numerical simulations but for graph-based applications such as data-discovery and deep learning. In public science and research, this share is still far below 10%. This is a result of the digital transformation process and topics associated with it. For example, in the energy sector, only fifteen years ago, HPC-use was associated with the simulation of combustion in coal- or oil- fired power plants and the simulation of incidents in nuclear power-plants (a ′classical′ numerical simulation). Today, the focus of the use of HPC in this sector has shifted towards topics such as matching the production of renewable, non-persistent and distributed sources of electrical energy with user requirements by exploiting variable loads, etc. These are typical machine-learning scenarios interacting in real time with cyber-physical systems.

We may see similar developments in the automotive industry in the near future. Trends such as electrical vehicles and autonomous driving will completely eliminate some classical numerical simulation scenarios, e.g. combustion (the most difficult HPC application in the automotive sector to date). On the other hand, data-analytics and machine-learning scenarios will grow in importance.

Generally speaking, the progress towards a digital society will move us to a higher abstraction layer, e.g. the focus will be the energy-supply rather than the power plant; the traffic, rather than the car itself, etc. The above pattern will have a significant impact on HPC. Not only new questions will arise (e.g. What is exa-scale in Machine-Learning or HPDA?), but also new research-fields (e.g. Cyber Physical Systems will attach HPC to the real-time and embedded world) have to be addressed.

4.3.3
Big Data and Extreme Computing (BDEC)

The initiative on Big Data and Extreme-scale Computing[74] (BDEC) is premised on the idea that we must begin to systematically map out and account for the ways in which the major issues associated with Big Data intersect with, impinge upon, and potentially change, the national (and international) plans that are now being laid for achieving exascale computing.

The goal of the BDEC workshops has been to develop an ICT planning document for science and engineering that articulates an analysis and vision of the conjoint evolution of data-intensive research and extreme scale computing. The findings and **technical** recommendations fall into three categories: global recommendations, recommendations for centralised facilities, recommendations for edge environments.

[74] www.exascale.org

4.3.3.1 Global recommendations

The major recommendation is to address the basic problem of the two paradigm splits – the software split and the data split. For this to be achieved there is a need for new standards that will govern the interoperability between data and compute. However, if we want a new distributed infrastructure to support science and engineering research in the era of Big Data, an infrastructure with the kind of openness, scalability, and flexible resource sharing that has characterised the legacy Internet paradigm, then we will have to define a new, common and open distributed services platform (DSP), one that offers programmable access to shared processing, storage and communication resources, and that can serve as a universal foundation for the component interoperability that novel services and applications will require. As the data revolution continues, such well-designed DSP infrastructure will be necessary to support such compute- and/or data-intensive work that many application areas will have to carry out between the ingress/egress to the Cloud (or Data/HPC centre) and the network edge. As the history of the Internet shows, the scientific community is, with appropriate public investment in basic research and development, uniquely positioned to create and develop the kind of DSP that the emerging era of extreme scale data and computing requires, building on the kind of open, consensus-driven approach that helped establish the Internet.

4.3.3.1.1 Recommendations for centralised facilities

1. Energy as an overarching challenge for sustainability. We can identify four steps towards energy minimisation: (1) reduce computational costs by using platforms well-matched to the stage within the scientific method; (2) reduce data-movement costs by co-location, compression and caching; (3) encourage re-use of calculations and data, by effective sharing, metadata and catalogues, a strategy that a provenance system supports well; and (4) reduce computing system entropy (workloads interference, system jitter, tail latency, and other noises) by on-demand isolation, noise-resistant priority, Cache QoS and novel uncertainty bounding techniques. The e-infrastructure itself has the task of taking care of energy minimisation as it has access to the required information - burdening domain scientists is not desirable, since it would divert them from their scientific goals.
2. Converging on a new hourglass. The "hourglass" represents the idea that an appropriately designed common interface can be implemented on an ever increasing variety of technology platforms (yielding a wide "lower bell"), while at the same time supporting an equally diverse and growing variety of applications (yielding a wide "upper bell"). The common interface, or "thin waist of the hourglass," is called the "spanning layer" because it bridges, through virtualisation, a heterogeneous set of resources that lie below it, but leaves the application and services above it free to evolve independently. This point clearly ties in with the global recommendation, above. Unfortunately, in seeking a new spanning layer to address the challenges of the big data era, the science cyberinfrastructure community finds itself in something of a dilemma. On one hand, at present there does seem to be at least one plausible and widely touted candidate for the new spanning layer operating-system-level virtualisation that supports software "containers". Certainly, converging on a common interface for containerisation would go a long way to achieving ecosystem convergence, and do so in way that requires something closer to evolutionary, as opposed to revolutionary changes to current modes of operation. Containerisation should thus be a very active area of research and experimentation across all the contexts that scientific cyberinfrastructure will have to address, including commercial clouds, HPC systems, and computing resources deployed in edge environments. At the same time, the fact that containers preserve legacy silos for storage, processing and communication at a low-level, and may therefore bring with them unexpected impediments to interoperable convergence, suggests that other ideas for a new spanning layer should also be aggressively pursued.
3. Data reduction as a fundamental pattern. The communication, analysis and storage of data from large scientific experiments will only be possible through aggressive data reduction, capable of shrinking datasets by one or more orders of magnitude. This is equally valid for data from sensors (see below). Although compression is critical to enable the evolution of many scientific domains to the next stage, the technology of scientific data compression and the understanding on how to use it are still in their infancy. Beyond the research on compression, scientists also need to understand how to use lossy compression. If the data needs to be decompressed, can we decompress it only partially to allow for pipelined decompression, reconstruction, and analytics? The same set of questions

applies to large-scale simulations: if we can avoid data sampling and decimation and compress the raw dataset by a factor of 100, can the following data analytics steps be performed on the compressed data?

4. Radically improved resource management. As HPC workflows start encompassing not just classical HPC applications, but Big Data, analytics, machine learning, and more, it becomes important to provide both the hardware and software support to run those workflows as seamlessly as possible. We define system management to be how a machine (or collection of machines) is controlled via system software to boot, execute workflows, and allow administrators or users to interact with and control the system. The roadmap to successful convergence requires freeing the user from the responsibility of managing the underlying machines themselves and enabling widespread use of these complex machines.

5. Software issues. As the new era of Big Data and extreme-scale computing continues to develop, it seems clear that both centralised systems (viz., HPC centres and commercial cloud systems) and decentralised systems (viz., any of the alternative designs for edge/fog infrastructure) will share many common software challenges and opportunities.

    a. Leverage HPC math libraries for HDA
    b. More efforts are needed for dense linear algebra standards
    c. New standards for share for shared memory parallel processing
    d. Interoperability between programming models and data formats

4.3.3.1.2
Recommendations for the edge ecosystem

1. A Common Distributed Services Platform for Data Logistics. There seem to be at least four, non-exclusive alternatives for interfacing HPC to this new paradigm in which no strong assumptions are made about where the data are: (1) data streaming, (2) in-transit processing, (3) processing at the edge of the distributed system, i.e. as close as possible to the data sources, and (4) logically centred cloud-like processing.

2. Cloud stream processing capabilities have not been designed with HPC in mind, and there is a need to examine the high performance aspects of their runtimes.

3. Commercial CDNs (content delivery network) are expensive to run and creates barriers to interoperation. To resolve this requires (1) a scalable approach to CDN implementation (i.e. suitably designed forms of storage and processing at the nodes of the distribution tree) and (2) the aggregate organisational will of the scientific community.

4. Software Libraries for Common Intermediate Processing Tasks. One common theme in the workflow descriptions is the amount of "intermediate" (or pre-) processing that data requires before the more substantial analysis and visualisation processes can occur. Some of these operations are generic enough so that a common set of software tools, appropriately layered and modularised, could be developed to serve the diverse purposes of a number of different communities at the same time.

4.3.4 HiPEAC's vision

HiPEAC[75]'s [HiPEAC] mission is to steer and increase the European research in the area of high-performance and embedded computing systems, and stimulate cooperation between a) academia and industry and b) computer architects and tool builders.

The HiPEAC Vision document is a deliverable of the Coordination and Support Action on High Performance and Embedded Architecture and Compilation that gathers over 450 leading European academic and industrial computing system researchers from nearly 320 institutions in one virtual centre of excellence of 1700 researchers.

The HiPEAC vision is published every two years since 2008 and tries to sketch the future landscape in the domains related to HiPEAC. The recent versions are organised following the following structure: societal evolution and its impact on ICT, the market trends, the advances and limits of technology, and the position of Europe in computing. From this information, a set of recommendations for the community is issued.

Here is a brief summary of the HiPEAC vision 2017[76], with a special focus on High Performance Computing.

[75] www.HiPEAC.net

[76] https://www.HiPEAC.net/publications/vision

The Information and Communication Technology (ICT) domain is evolving rapidly and new challenges are always ahead of us.

The computer is disappearing from view, yet settling in the very fabric of everyday life. It takes on new forms, not only those of servers, PCs, supercomputers, smartphones and tablets, but also as cars, smart meters, thermostats, and so on. They communicate with their users not only through keyboards and alphanumeric display screens, but also using voice, sound, pictures and video, closely resembling human interaction.

The function of the computer is shifting, from computational tasks providing answers to numerical problems, to humans and computers working together (what we call the beginning of the Centaur Era).

But more is to come: interactions with computers will be augmented by virtual reality, modelled as interactions between humans, and all this made possible by the use of Artificial Intelligence-based techniques. This will not only change how we interact with machines, but it will also redefine how we instruct a machine what to do: less programming and more learning.

This could also have a drastic impact on research: Microsoft states that we are currently using the fourth paradigm of scientific discovery. The first three paradigms were experimental (empirical description of phenomenon), theoretical (discovery of laws, models, etc. able to predict results) and, more recently, computational science (computer simulations). The fourth paradigm of scientific discovery is the analysis of massive data sets, enabled, e.g. by data capture, curation, mining and analytics techniques and thus permitting new scientific discoveries. In the fourth paradigm, computers are used to extract information from raw data, but it is still humans who perform analyses of the information and make the scientific discovery. We believe that within the next decade there will be a fifth paradigm, in which computers will be not only extracting information from data, but will also formulate a hypothesis, invent new simulations or make new formal proofs and finally make themselves scientific discoveries without human intervention. Precursors can be seen with formal provers, data analytics systems, etc.

Computing systems not only observe, but increasingly interact with the physical world by controlling it. Such systems are called Cyber-Physical Systems (CPS). The most visible development in early 2017 is the Advanced Driver Assisting Systems (ADAS), which will evolve into autonomous driving cars. This will drive a shift from security to safety and trustability: because of the direct control of physical devices, a malfunction of a computer, due to a programming error, hardware failure or a hacker, could have lethal consequences. Humans need to trust the machines, not only by behaving in a correct and predictable way, by having understandable decisions but also by keeping sensitive information about the human confidential. Therefore, enforcement of transparency, security and privacy are of paramount importance.

These developments pose large challenges to the HPC and to the HiPEAC communities.

Figure 12.
The main HPC challenges of the HiPEAC communities (courtesy of HiPeac).

- HPC at the edge: safety, security, and economical reasons (communication bandwidth) will drive high performance computing at the edge: supercomputers from previous generations will become embedded systems in the next generations

- HPC in the loop: simulation of complex systems will become more and more demanding and in close interaction with the real world and with users. Simulations that were previously independent will be integrated in global frameworks and will need to be interoperable. Simulations (e.g. of industrial

processes) will have more "real-time" requirements for being able to be used to forecast behaviours before they arrive. In systems design, to master complexity we need methodologies that enable composability and interoperability of components, and we very well may need to add AI-based techniques and tools to help mastering this complexity

- Human in the loop and interactive HPC: users will be able to observe and change simulation parameters dynamically, during the course of the simulations, not waiting for their end

This interactivity will also drive the development of approaches presenting a large amount of data in an understandable way for human to decide.

- Convergence of systems supporting simulations, data analytics and Artificial Intelligence.

Deep Learning already requires exaflops for the learning phase, and future systems will have both to run complex simulations, the capability to process large amount of data (data analytics) - generated by the world or by simulation-, and artificial intelligence capabilities for interpreting and taking decisions.

Other challenges are driven by technology itself:

Energy efficiency of computing systems remains a major challenge for the coming years, and not only to decrease their environmental footprint: without a significant improvement in energy efficiency, Exaflops computers will not be economically viable and the myriad of small (battery-powered) computing devices will not be successful due to their lack of autonomy. It may well require breaking away from the traditional Von Neumann architecture and rethinking device technologies. It is also the right time to revisit the assumptions that drove the semiconductor and computer industry for decades, and to challenge its explicit and implicit assumptions in order to open new tracks and new approaches and to eventually not reinvent computing, but completely rethink the basic concepts of computing. Reducing or eliminating the transfer of data (computing in memory) is one example of a potentially interesting track.

As the cost per transistor is no longer decreasing, and even appears to be rising, we might see diversified tracks for using silicon technology: many designs will not use the latest technology node, but the more mature (and cheaper) one. Only high performance systems will require a very expensive and aggressive state of the art technology node. Diversity in terms of coprocessors and domain specific accelerators will also be key to ensure the highest efficiency of the global system, at the cost of increasing (software) complexity.

For all these challenges, one very important requirement is: we can only achieve solutions if we adopt a holistic approach in the development of high performance systems, beyond current co-design, where all disciplines come together and are regarded as first class citizens.

The complexity of the new systems will be so high that human designers will only be able to master it with the help of computers using AI-based techniques. Innovative approaches will be required to ensure that the systems will do what they are supposed to do, both at the functional and at the non-functional level (e.g. energy requirement, timing requirement and reliability). We need to develop design techniques that go beyond predictability by design and allow the building of reliable systems from unreliable parts.

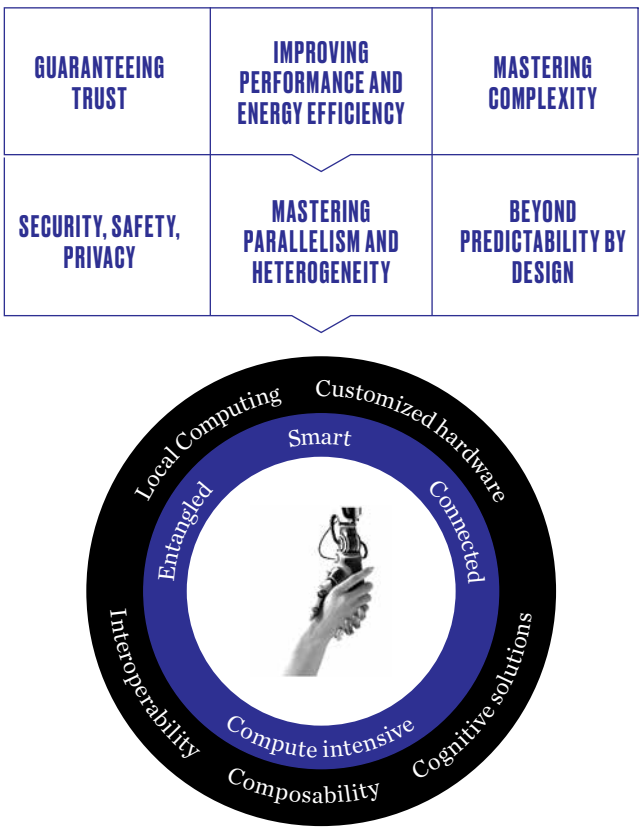


Figure 13.
The general insights of the HiPEAC Vision 2017.

5.

NEW TRENDS IN HPC CHALLENGES, USE AND TECHNOLOGY

To keep up with the ever demanding HPC user needs, the performance of HPC systems at all scales must continue to grow faster than the regular progress of computing technology. More powerful processing units, new memories, storage, networking technologies and larger systems in size require changing the **architecture of HPC systems**.

Most significantly, the recent use of HPC accelerators (e.g. GPUs, Many-core CPUs) has resulted in a significant performance boost for some applications. To enlarge the application scope of these HPC accelerators they must be better integrated in the architecture of the HPC nodes or at the system level. To deliver performance, these HPC processing units must access data with a much higher bandwidth than is available with today's DRAM memory technology. This is possible with the introduction of high-bandwidth memory in addition to or in lieu of DRAM. To overcome the limited capacity of these fast memories and to fully leverage their potential, a complete re-engineering and re-architecting of HPC applications might be required.

New non-volatile memory technologies (NVRAM[77], also referred to as "SCM-storage class memory") appear to be on the near horizon. It is expected that they will offer a much larger capacity than current DRAM, a good fraction of DRAM bandwidth and similar endurance. This opens interesting opportunities for the design of HPC systems. Long-term, the new NVRAM could replace DRAM altogether in compute nodes and become the base for ultrafast storage at the same time. These are technologies which would greatly improve the ability to save application state in checkpoints and restart runs which have failed from them.

HPC systems are highly parallel. The many thousands of nodes available in an HPC system must be tightly coupled by a low latency network which also integrates storage. The HPC system network must scale with the number and performance of compute nodes and storage devices, requiring more bandwidth but also cutting latencies. The HPC network should also provide sufficient functionality to access to the whole system resources in a simple manner.

Virtualisation is making its way into the HPC system design, it allows a more flexible usage of HPC systems which is valuable as long as performance impact remain low; it would increase system security resiliency and flexibility. Check-Point/Restart mechanism and exclusion of defective nodes can take benefit of it. Network Virtualisation will be an important part of this.
Finally, new High Performance Data Analysis (HPDA) applications are emerging alongside with HPC. The two fields interact with each other and are identifying the common ground regarding requirements and technology.

System software plays a key role in ensuring that future HPC systems can be used in an efficient way: it provides critical abstractions for programming environments and applications, makes new system features and functions accessible, and manages the increasingly complex and diverse set of resources that make up a large HPC system.

- System and node architectures will become more complex as exemplified by the discussion in the "HPC System Architecture" section of this document. Proposals should target mechanisms for adaptive and dynamic scheduling, management and use of heterogeneous system components

[77] Non-volatile random-access memory

to achieve the energy efficiency and resiliency objectives whilst meeting application performance requirements.

The key themes identified in previous versions of the SRA retain their importance. In addition, the increasing amounts of data generated and processed by HPC workloads, as well as the rise of data analytics applications create the need for highly performant and efficient, large storage that is well integrated into future HPC and HPDA systems. Upcoming storage-class memory technology is poised to meet these needs: advanced 3D flash and Intel/Micron's 3D XPoint™ technology, for instance, promise to deliver high capacities, data persistence and significant access performance increases, plus cache-line or Byte addressability.

To support this storage evolution new drivers and libraries are required which make novel block or byte access to data available, and higher-level runtime system and I/O components have to be adapted to leverage these. In addition, innovative data-oriented programming models can emerge, and the new storage will be integrated into efficient PGAS models.

The increased use of heterogeneity (e.g. FPGAs and other reconfigurable computing systems) both within nodes and in the system will significantly impact the system and run time software layers as well as the integration of future interconnects. We foresee two vectors:

1. Enable simple and efficient use of all heterogeneous compute components, for instance through specialised micro-kernels or containers

2. Support optimal use of the system-wide resources through intelligent and flexible resource aggregation and orchestration mechanisms

Currently, there are a number of (competing) standardisation efforts on the integration of accelerators within the node and on remote storage, with CCIX, Gen-Z and OpenCAPI as examples. The future convergence of these standards would support sustainable development and uptake of heterogeneous HPC and HPDA systems, provided that said standards do not unduly impact the opportunities in energy reduction and performance increases provided by close physical integration.

- The evolution of interconnects towards higher bandwidths and lower latencies will continue, with the Ethernet family of fabrics keeping the pressure up on improving bandwidths. In addition, efficient support for RDMA, peer-to-peer communication and collective communication patterns are important drivers.

Capabilities for adaptive routing and dynamic congestion avoidance and QoS guarantees (e.g. of injection rates and bandwidths) will become more and more critical to ensure efficient use of systems towards the Exascale size range.

- Data movement between devices on a node and between nodes accounts for a substantial and increasing part of energy use in modern systems. Therefore, data transfers at all system levels have to be minimised, by using advanced techniques including data-aware scheduling, in-situ or near-memory data processing, and potentially the use of communication-avoiding algorithms. Data transfers that cannot be avoided have to be optimised through highly efficient communication and remote data access mechanisms, and appropriate data protection has to be provided.

With the rapid evolution in storage-class memory and its tighter integration into future systems, mechanisms for efficient storage access over a fabric will become important; one example here is the NVMe over Fabric effort.

- New application domains such as data analytics and AI-driven machine learning (today, mainly Deep Learning) are quickly gaining ground. Since the earlier SRA, HPC workflows that combine simulation and data analytics have emerged (such as in brain simulation, high-energy physics, numerical weather prediction, astronomy, etc.), and an increasing set of HPC operators are actively extending the range of customers and workloads they support to include data analytics. Visionary use cases such as fully autonomous driving or precision medicine will require a tight integration of HPC, data analytics and AI/ML, while promising large market opportunities.

These new domains mean new software stacks and new workflows which can benefit from the huge available parallelism, performance, and scalability of large supercomputers. It will be critical to identify the specific requirements of these stacks and workflows and to guide the development of system architecture, system SW and programming models accordingly.

As an example, today's Deep Learning requires fast execution of dense linear algebra operations on oddly shaped matrices, and fast distributed gradient descent. The former greatly

influences CPU and node design, while the latter impacts the inter- and intra-node fabrics and communication SW stacks.

- The new usage scenarios require new, dynamic and flexible execution models and support for complex, multi-step workflows. Research should target efficient orchestration of such workflows, including co-allocation of work, sharing of compute and data resources between steps, and compute-near data with the objective of optimising end-to-end performance, system throughput and energy efficiency. In addition, efficient integration of virtualisation or container approaches that would improve ease of use, efficiency and resilience of systems can be considered here. Work proposed should treat data as a first class resource and take into account needs for extending resource management, orchestration and scheduling functionalities.

New system architecture principles do implement the concept of resource disaggregation and software-defined infrastructures, with Facebook's Open Compute Projects and Intel's Rackscale Design Architecture as examples. Such approaches can bring efficiency and TCO[78] improvements for HPC applications with time-varying resource requirements, yet it is the new combined HPC and analytics use cases that will profit most. Research in this area can have a substantial impact on the design of systems that target both HPC and data analytics. Whilst the long term focus of HPC hardware and algorithm development is the minimisation of execution time for individual applications/jobs, the convergence of previously discussed hardware trends and full-workflow considerations motivates an additional focus on maximising throughput of scientific workflows as well as minimising the time of the computational simulation parts of the workflows

Standardisation of the system software architecture and APIs[79] will no doubt encourage adoption of the new breed of heterogeneous HPC/HPDA platforms across a wide range of systems and markets. This will require a coordinated approach between programming models and runtime systems, and will take significant time and effort. The development and integration of specialised containerised execution environments may provide a shorter and less risky way towards technology take-up.

Several key, high-level **programming environment** themes identified in previous versions of the SRA retain their high significance, but require advancement. The overriding theme is effectively supporting applications in their transition to exascale.

- Enabling effective application development and deployment at extreme scales requires high-productivity programming environments which: reduce programming complexity; separate core algorithmic issues from implementation and optimisation concerns; facilitate the expression of algorithmic parallelism and asynchronicity in applications; ensure code maintainability and portability across existing and possible future architectures and systems. Portability is understood to be in the first instance functional portability but where possible also performance portability, potentially by facilitating optimisation without the requirement for large-scale impact on the application. At the same time, the need to support full-scale production applications and the migration of legacy software suites – in both cases ensuring efficiency and scalability- is essential.

- The development of programming models and associated runtime systems and compilers requires co-design both with the other SRA domains (the efficient interaction with the system software level includes the need for direct support functions) and also with highly relevant applications. Programming models need to take information from the application's computing system use & requirements into account and the scalability of the realisation of programming models and software tools at all levels is a key concern.

- There should be strong interoperability throughout the programming environment, including compiler tools and runtime systems, debuggers and performance tools, linking information to the programming model and source code.

An aspect with growing importance, and one linked to the interoperability topic mentioned above, is to establish the acceptance and adoption of the programming models and application programming interfaces (APIs) in industrial and scientific production codes. This requires an emphasis on long-term standardisation of programming models, APIs and the related runtime system functions and tools.

In the area of modelling and simulation (well-established industrial and scientific computing), the development of the programming environment towards exascale benefits from a relatively long past history of HPC use. For high-performance data analytics ("HPDA", used here to include Big Data and deep learning/machine learning applications) the situation

[78] Total Cost of Ownership

[79] Application Programming Interface

is rather different, with the trend of very rapid development of programming frameworks and languages, tools and software systems. It is however recognised that a convergence of HPC and HPDA offers great opportunities and that there is a significant potential in the identification of commonalities in the software stacks and in the possible provision of "cross-area" programming environment components. Ensuring that HPDA frameworks are able to make effective use of the key components of HPC-like systems would be an important first step. The investigation into these possibilities is a new research challenge.

Traditional workflow management systems have distinct phases for computation/simulation and data analysis. Dynamic workflow systems are required in order to support the convergence of HPC and HPDA, in particular to support in situ data analysis and visualisation. Such systems will couple simulation, databases, data analytics and visualisation, and the results from intermediate data analysis steps should be able to trigger detailed simulation steps.

The cost of the **electrical energy** driving the Information and Communication Technology (ICT) is an important consideration. The 2010 estimates put the ICT at 3% of the overall carbon footprint, ahead of the airline industry.[80] Modern large-scale datacentres are already multiple of tens of MWs, on par with estimates for the future Exascale HPC sites. Therefore, computing is among heavy consumers of electricity and subject of sustainability considerations with high societal impact.

The application advance defines the efficiency metric as follows: *Energy Efficiency = Useful Work/Energy*

The definition of "Useful Work" is deferred to the discussion of the applications.

Another source of the energy efficiency increase is related to the resiliency. Energy is lost if computation has to be repeated due to failures and therefore we consider techniques to recover from the failures in the machine.

For the HPC sector, the key contributors to electricity consumption are the computing, communication, storage systems and the infrastructure including the cooling and the electrical subsystems. Renewable and energy neutral ways of organizing the data centres is gaining popularity in the US and we expect it to become a major driver in European Data Centres as well.

To that end it is expected that the (pre-)Exascale IT sites, while initially incorporating direct liquid cooling for the computing clusters proper, will further evolve toward (direct) liquid cooling without use of air for the heat transfer for the data centre overall.[81] Cooling with temperatures of the liquid above 45°C opens the possibility for the "free cooling" in all European countries and avoids the energy cost of water refrigeration.

The communication system is an essential component of the data centre and a significant energy consumer[82]. Typical data centre network may consume up to 10% of the delivered power. Therefore, we expect a trend towards faster, full optical networks with fewer interfaces to save the energy.

The management software will balance the energy spent in the datacentre with its computational output taking into account all aspects of sustainability and local policy. For the aspect of the Energy Efficiency we single out the monitoring, the ways of optimisation of the energy spend for specific computational result as well as the power policy at the facility. For this purpose a comprehensive sensor network will be developed as part of the Data Centre facility and the arising Big Data problem of handling that information to manage the machine will be solved.

In the context of technological issues, we consider the devices at the heart of the computation: processors and memories organised in nodes and servers. These spend most of the power delivered to the data centre and provide biggest opportunity for the energy efficiency increase. The reduction is along two lines: (a) the reduction of the electrical energy spent while obtaining the computational results. This is possible by increasing the efficiency of data manipulation and computation itself (may be measured in Flops/watt or any other application dependent metric); (b) efficient extraction of the heat produced during the computation.

The new packaging technology puts silicon chips in close proximity on a single interposer (2.5D integration). More advanced packaging puts them on top of each other (3D integration), connecting them by the Through Silicon Vias (TSV). Both of these technologies reduce the energy needed for signalling as shorter connections need less switching power. However, the cooling complexity must be mastered when using this approach. These technological aspects need to be demonstrated to support the European Low-power processor initiative, irrespective the roadmap for the final processor.

[80] L. Smarr, "Project greenlight: Optimizing cyber-infrastructure for a carbon-constrained world," Computer, vol. 43, no. 1, pp. 22–27, 2010.

[81] The DEEP-ER ref. **http://juser.fz-juelich.de/record/202677**, From 2013-10-01 to 2017-03-31, FP7-ICT

[82] For the global view and the societal impacts, see **http://www.iea.org/publications/freepublications/publication/more-data-less-energy.html**

The overall computational process must be organised such as the outcome of the calculation should be proportional to the energy spent. Keeping this 'Proportional Computing" philosophy in mind it is necessary to characterise application advance to optimise the usage of resources in general and spent energy in particular. Power is instantaneous value of electricity, while energy is an integral of that quantity over time. The application can be optimised for power and/or energy separately and the minimisation functions should be formulated for both depending on the metric chosen. We expect the collection of multiple application optimisation strategies should extend the policy to the computer centre by providing this information to management software.

In the area of **resiliency**, preserving data consistency in case of faults is an important topic in HPC and a way to optimise energy usage as failed applications may need to be repeated. Individual hardware components can fail causing software running on them to fail as well. System software would take down the system if it experiences an unrecoverable error to preserve data consistency. At this point, the machine (or component) must be restarted to resume the service from a well-defined state.

No deterministic failure prediction algorithm is known. However, collecting sensor data and Machine Learning (ML) on this sensor data yields good results.[83] We expect that the (pre-) Exascale will incorporate sufficient sensors for the failure prediction and monitoring.

An advance in relation to the current state of the art will be the implementation of the checkpoint mechanism that does not need to be coordinated and can be done per node when the computation reaches a well-defined state. The non-volatile memories may be available for the checkpoints. It is clear that the system software will incorporate failure mitigation techniques and may provide feedback on the hardware based resiliency techniques. On this note, the compiler assisted fault tolerance may bridge the separation between the hardware-only and software-only recovery techniques.[84]

The Algorithm Based Fault Tolerance (ABFT) may also use fault detection and recovery from within the application. This requires appropriate data encoding, algorithm to operate on the encoded data and the distribution of the computation steps in the algorithm among (redundant) computational units. The ABFT is expected to be able to detect and to recover from the silent data corruption.

In the context of **balancing Compute and I/O performance**, Compute and Disk drive performance continues to lag behind compute performance and capabilities (including higher thread counts and heterogeneity). Disk drive capacities and aerial densities are growing much faster than data access performance. There is hence a continued push to close the compute and I/O performance gap through intermediate tiers such as Flash as seen by industry initiatives, requirements from big science communities and national labs. The availability and cost of Flash technology in the storage ecosystem has enabled the development of Burst buffer technologies in HPC – with the Flash tiers acting as performance intermediaries between the compute and storage subsystems. However, there are new devices appearing on the horizon (in spite of delays in vendor roadmaps) whose usage within the HPC I/O stack is not very clear. Study of these NVRAM technologies and their usage within HPC will be the next major trend especially with samples of 3DxPoint technology now becoming available (Note: In the previous SRA2 update we had indicated that it was still not very clear when these NVRAM technologies would become available) and the continued anticipation and expectation of more NVRAM technologies.

A new trend that is developing as a subsequent outcome of the above, and in parallel with developing trends such as memory centric computing is the usage of persistent storage just like memory to provide a uniform and simplified interface for programming environments. These will provide for possibilities such as flat global addressing of data independent of whether it resides in memory or persistent storage in the HPC I/O stack blurring the lines between them. How the applications will truly exploit such a trend needs to be further studied.

Convergence of storage infrastructures for Big Data Analytics and High Performance Computing is continuing to happen more actively than at the time of the SRA2 update. There is a continued need to manage and process more amounts of data from instruments and sensors to make them part of the work flows. This in turn requires the continued development of methods such as process and task offload to the storage system, and storage systems technology evolution to provide ingrained compute capability as part of the storage system. Considering that energy, or more precisely, scientific throughout for a given energy envelope, is one of the critical aspects of scaling, avoiding data movements will provide for drastic energy savings considering that moving data to compute engenders orders of magnitude more energy than doing "in-place" computations wherever data originally resides. This of course will

[83] Doug Turnbull, Veil Alldrin, "Failure Prediction in Hardware Systems", http://cseweb.ucsd.edu/~dturnbul/Papers/ServerPrediction.pdf

[84] T. Herault, Y. Robert, ed. "Fault-Tolerance techniques for high-Performance Computing", Springer 2015.

be aided by the support of smart infrastructure components that can turn themselves off or on based on the need to transfer data (e.g. networking and switching fabric) or process data in place(e.g. disk drives). Classical energy saving techniques for disk drive arrays that had been previously proposed for the enterprise (e.g. MAID, or Massive Array of Idle Disks) may find a niche in extreme scale HPC again. It is indeed important to look at energy consumption holistically from the perspective of whole workflows or large ensembles.

Object storage technologies will continue to be at the forefront of storage software innovation in HPC as a continuation of the trend we had indicated in the SRA2 update. Parallel file systems, due to the Posix access model, will reach fundamental bottlenecks in terms of performance and their scalability – given that they were designed for an era when the assumptions for data volumes, performance requirements and computing environments were very different. There is now an even greater push from the open source community (than in 2013 – 2015) as well as by commercial entities to have the adoption of Object storage within HPC.

AI and deep learning workloads are now the new class of workloads that need to be studied with regards to their data management, storage and I/O needs as they are starting to show increasing need for the usage of HPC resources for problem solving. Increasing amounts of memory within compute nodes and need for faster interconnection networks will become the norm for such workloads. This is a continuation of the trend we had indicated in the SRA2 update for Big Data Applications in general. However with the memory footprint per core reducing, there will be greater focus on the usage of Non Volatile Memories for AI and deep learning workloads. This indicates greater opportunity for data access optimisation and data management employing the storage software stack traditionally reserved to the confines of persistent storage.

Federation of data across the cloud with the emergence of HPC in the cloud and HPC at the edge, and new Big Science scientific initiatives where extreme volumes of data are generated at different geographies and need to be managed and processed holistically as part of big scientific workflows, is the next big trend. We had already indicated data federation from highly distributed sources in the SRA2 update. HPC in the cloud has also provided greater impetus for the study of storage, data access and I/O optimisation in virtual machine environments within HPC, which has typically been the realm of enterprise storage. This also brings about the problem of data security that was not a big consideration earlier.

Having flexible and feature rich storage APIs to deal with various types of data sets, providing storage system feature extension capability by third parties, big data analytics inclusion and including evolving programming models continue to be of very active interest. The focus should be on portability (functionality and performance) and the evolution of well-defined standards.

Data Management software needs to evolve to keep up with the data volumes and varieties, as we get to higher scales. Data Management deals with aspects such as Information lifecycle management, tracking provenance, etc. We should provide well suited tools to help communities to define and implement their data management plans.

Faults will continue to be a norm rather than an exception as we indicated early on in the SRA2 update. There is an increased need to deal with constantly occurring faults within the system (including storage software) without causing any downtime to applications. The storage system is becoming even more complex with more tiers and more software ecosystem components (and compute components within the storage system). Reliability and resiliency of such systems will be extremely critical. As a corollary, there is a need to provide much deeper visibility into system performance through exploiting aspects such as deeper and richer storage systems telemetry data.

Quality of service mechanisms where the executed workload receives performance guarantees from the storage and I/O subsystem - need to receive far better coverage in the HPC community though there has been only incidental works in this direction. One associated aspect of this is that the storage system needs to be better integrated with batch schedulers, wherein the schedulers can provide hints on data usage, etc. that aids in placing data at the right place in the storage hierarchy at the right time before the appropriate pieces of the workflow are executed.

The storage and I/O subsystem needs to specially optimised for low power processing environments and exploit some of the features and capabilities they offer. This is important considering the focus on overall energy reduction, and having a storage system that is ready and able to work under these constraints.

Dealing with **extreme data** is a situation that manifests itself with fast increasing frequency of HPC systems. The root causes stem from two main categories:

Traditional HPC simulations (e.g. fluid dynamics, computational chemistry and physics, cosmology among others) have been benefitting from the tremendous increases in computational capacity of HPC systems and are now using models of unprecedented realism and accuracy. These models represent the real world using trillions of discrete degrees of freedom, which require huge RAM and scale out systems. These simulations generate enormous amounts of output data. On the other hand, researchers need to apply advanced and highly complex analytics and processing (including visualisation) on this data, which simply means that off-loading to remote platforms is simply not an option. Thus, data analytics needs to take place in-situ, and perhaps in synchrony with tightly coupled synergistic computing platforms (e.g. visualisation engines).

On the other hand, new applications arise as potential HPC clients. Big Data applications, in which data is not generated from some sort of model but rather is collected, accumulated or even streamed, and comes with computational complexity that already sets the computational needs to the petascale or even to the exascale region, even after local pre-processing. Big Data systems, that have been primarily developed for scale out to distributed, non-reliable, resources, are simply too coarse in efficiency and cost effectively to cope with such computational complexity. Thus, HPC solutions, with lots of memory and very fast networks start to appear very appealing and have already started to be used as tools by Big Data users. Perhaps the most prominent example is Deep Learning, which relies on large number of accelerators in order to speed up training to an acceptable time to solution. Moreover, Big Data systems have already started to be influenced from HPC architectures and practices (e.g. multi-threading, parallelism, memory modelling etc.). On the other hand, it is also clear that data will be highly distributed as it originates from distributed sources. Thus, synergy of Big Data systems and tightly coupled HPC systems is foreseen in forming a larger, hybrid, computing resource that combines local and global processing to serve the needs of the new kind of Big Data applications. This trend however, means that the classic mode of use of HPC systems will need to be adapted to take into account key aspects of Big Data requirements such as security and privacy.

Advances in **mathematical methods and algorithms** will be essential in order to produce robust applications that can leverage future high-performance exascale architectures and to reach the goal of improving energy efficiency by two orders of magnitude. Significant efforts in this area are required to allow applications to become even more parallel, scalable and robust, and to optimise for data locality on architectures with deepening and heterogeneous memory hierarchies. New challenges arising from emerging application areas that require both, high-performance computing resources, as well as the ability to manage and process extreme amounts of data should be addressed. The impacts of research on mathematical methods will not be confined to applications, but will equally influence the design of future exascale system software such as compilers, communication libraries and programming environments. Notably, the area of optimisation and scheduling will benefit from new mathematically motivated approaches. The interaction with exascale programming environments will be important, as those environments will provide the platform through which new mathematical methods and algorithms will be realised in software applications.

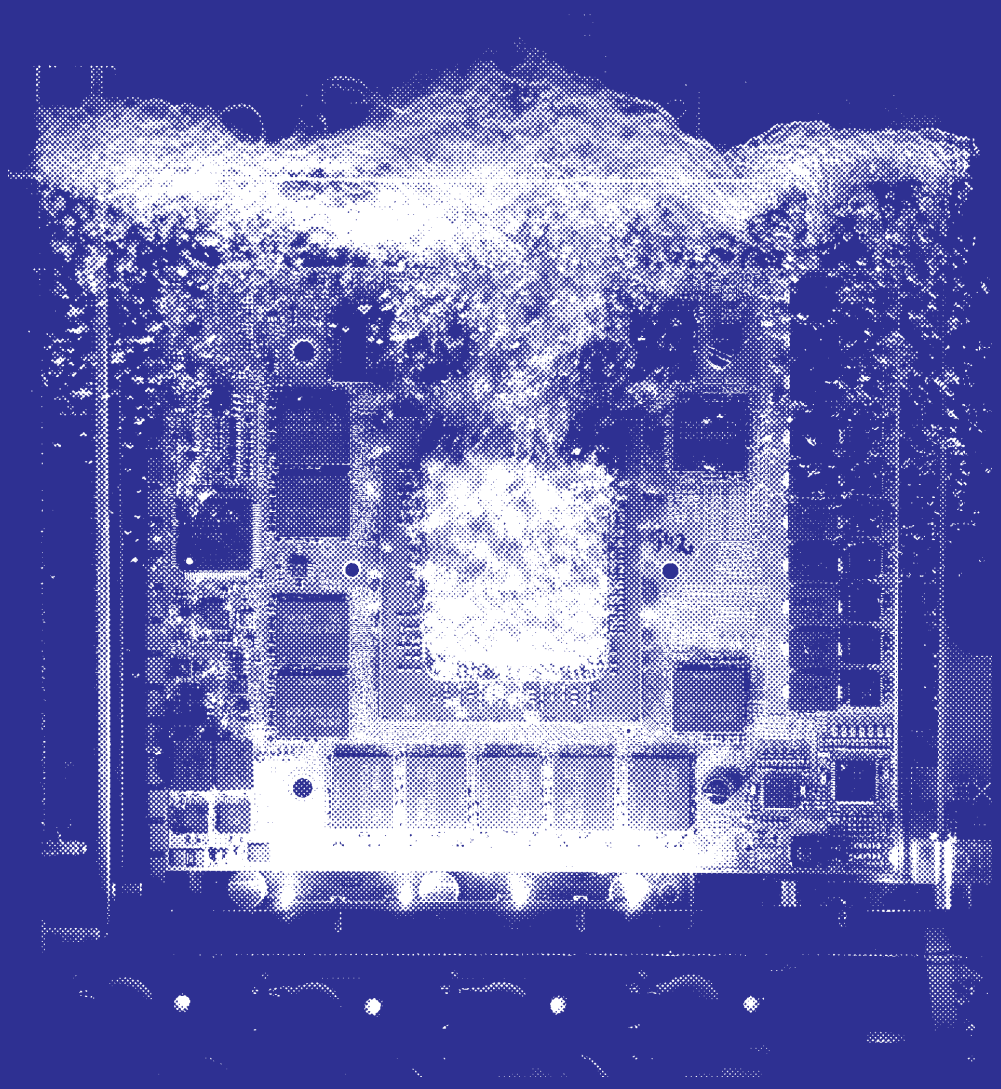
Mathematical methods and algorithms is a relatively new area addressed within the ETP4HPC's SRA. The agenda will focus on research required to achieve extreme scale performance, and in particular, take into account the impact on the commercial stakeholders that include:

- Users of HPC resources that benefit from the development of new methods for solving numerical and extreme data challenges, in particular, in the area of industrial and engineering applications as well as emerging Big Data applications

- HPC hardware system vendors who look for algorithms that enable most efficient exploitation of their solutions

- Software solution and service providers (e.g. ISVs), for which new and robust mathematical methods and algorithms are crucial in providing more competitive software solutions and to enable SaaS services, e.g. based on HPC in Cloud concepts

6.



TECHNICAL RESEARCH PRIORITIES

The following Sections (6.1-6.7) describe the areas of European HPC technology in detail. Section 6 lists the milestones to be accomplished. These milestones are referred to in the text of Section 6 using the following convention, e.g.: **[M - AREA CODE – NUMBER]**.

6.1 HPC SYSTEM ARCHITECTURE AND COMPONENTS

To meet the ever demanding requirements for performance, HPC systems must keep evolving. Simply following technology evolution is not enough as improvements are too slow in meeting exascale targets. In particular, HPC systems will feature much more powerful nodes using HPC processing units, faster/larger memory and storage devices, and also better interconnect. Energy efficiency is the main roadblock towards exascale as a significant increase of performance is required, whilst only a small increase of the power budget is affordable.

6.1.1 Compute nodes – HPC processing units

There are two trends in the recently deployed HPC systems: HPC accelerators such as GPUs attached to generic processors and HPC standalone manycores. Both of these units are highly parallel and use hundreds/thousands of threads to deliver the performance. Setting a higher number of threads with a lower single thread performance within an HPC accelerator might improve power efficiency but as a result the applications scalability would have to increase significantly higher. As an illustration, the number one of the Top500 HPC systems integrates 10 million cores. In all solutions, cores run slower than those in generic CPUs, and thus, will not be usable for single thread performance sensitive applications.

To improve their efficiency, these HPC processing units when they are used as accelerators of generic processors, must be better integrated into the system architecture by improving latency, bandwidth and energy efficiency of their interface to access main memory and network (PCIe Gen4, CCIX, OpenCAPI, Gen-Z...) **[M-ARCH-1]**.

Another path of improvement of the HPC processing unit comes from making resources more dynamic for adapting hardware behaviour to application performance needs as voltage and frequency settings and enabling/disabling functional units.

Even with these improvements, the architecture might remain too complex for large HPC applications. Programming Environment should provide a standard interface to hide this complexity; at the hardware level the appropriate features should be available to support a hybrid approach as MPI+X (where X could be OpenMP or others) or as PGAS +X.

6.1.2 Data Access — HW components

Although CPUs and HPC accelerators have improved drastically, the peak performance of the compute nodes and the delivered performance at the applications level have been lagging. The growing gap between theoretical and delivered performance is directly connected to slow improvement in memory speeds; this is often referred to as the "memory wall".
To overcome this limitation, 3D and 2.5D technologies allow a new level of integration by putting several dies into a same package; the larger I/O count interconnecting dies that they offer, enables the integration fast memory (GDDR, MC-DRAM, HBM...) and HPC accelerators in the same package.

These solutions offers much larger bandwidth at same latency level but showing a smaller capacity than today DRAM memory modules since package size is limited, and thus, if external memory is still required, an extra level in the memory hierarchy is added. To add complexity, these high bandwidth memories must be either addressed explicitly by the programmer and/or run-time systems or used as a functionally transparent new caching layer. Even if the second approach provide a much simpler programming solution, quantifying its performance benefit remains a research activity since it depends on application characteristics (memory footprint, memory access patterns, etc.)

In general, delivering improvements in bandwidth and latency can have major impacts on code efficiency[85]. Thus, the challenge is to find the right balance between future memory characteristics (bandwidth, latency, size, power consumption, integration and cost) and usage (explicit data placement, automated placement or even caching).

Upcoming Non-Volatile Memory (NVRAM) technologies are opening new opportunities for HPC systems. New NVRAM will feature much larger byte-addressable capacities as DRAM: hundreds of GBs vs tens of GBs. Their performance (read or write BW) will be much better than current FLASH based NVRAM, approaching DRAM levels. Furthermore, their endurance should be comparable with DRAM at least in combination with some hidden wear-levelling technology. They could be used in HPC systems, both as main memory and ultra-fast IO, and would completely change the system programming model which distinguishes memory and storage (files), see also Multi-tier storage in Balance Compute, I/O and Storage **[M-ARCH-2]**.

As stated above the downside of adding more memory hierarchies is the increased complexity and more limited portability of applications. This is even more valid if the classic separation between "memory" and "storage" with different access semantics is being retained. A more favourable alternative could be a pure memory semantic fabric that handles all communication as memory operations such as load/store, put/get and atomic operations typically used by a processor. Such an alternative approach will become more attractive with full photonic interconnects and photonic switches and/

[85] http://memsys.io/wp-content/uploads/2015/09/p31-radulovic.pdf

or crossbars that provide direct optical links from processors to the full spectrum of memory and storage devices. Efforts to create open standards for this field are taking place, e.g. in the Gen-Z (http://genzconsortium.org/) and OpenCAPI (http://opencapi.org/) consortia.

6.1.3 HPC Systems Interconnect

HPC systems are composed of a large number of nodes, from 100 for a departmental system to 10,000s for the Top10 systems. The application performance relies on parallelism and depends directly on the efficiency of the interconnect unifying the compute nodes into a single system. The HPC system interconnection network must scale together with the compute nodes and the storage performance. The HPC networks bandwidth is planned to grow from a 100 Gb/s today, to 200, then 400 Gb/s in the coming years thanks to the development of new generations of SerDes circuits (Serialiser/Deserialiser converts data between serial data and parallel interfaces in each direction) **[M-ARCH-3]**.

With such transmission rates, the possible range for electrical connections will remain limited to less than two meters and optical links will become prominent. Underlying network technology improvements are expected in areas such as silicon photonics and photonic switching, which should enable scale performance whilst keeping consumption under control **[M-ARCH-4]**.

Independent of the link bandwidth increase just mentioned, Network efficiency (latency, message rate) is another topic of interest. Reducing communication latencies requires a better interconnection of Network components with compute and memories by using more efficient connection than current PCIe (new generation of PCIe, new coherent link, etc.). Other improvements will come from the integration of network accelerators within either the network controller or fabric switch such as collective accelerators, MPI accelerators, etc.

Improving the network fabric in terms of quality of service is an important a target; Network hardware components must provide a good support for virtualisation (virtual networks, service classes...) for increasing traffic isolation capability that can be by traffic type, size and users; routing algorithms and resource scheduling can also bring benefits in isolating traffic (IO vs compute, applications among themselves, etc.) **[M-ARCH-5]**.

An optimised hardware support in term of performance of a direct access to the whole system memory would enable new ways to program parallel applications. In particular, the evolution of today's PGAS (Partitioned Global Address Space) programming languages will simplify the implementations of HPC applications **[M-ARCH-6]**.

As stated in 1.1.2 in more detail a memory semantic fabric – independently of the chosen network topology - would ease efficient implementations of programming paradigms that rely on global address spaces.

6.1.4 Global Energy efficiency

Moving to exascale implies more powerful compute nodes in a larger count but it has to take into account the electricity energy bill of hosting sites that must remain affordable. As a consequence, energy efficiency is the major issue for the design of exascale HPC systems design. Although, this theme is an essential motivation in all system components development (compute processors, memory, storage, interconnect, power system, cooling and power delivery, etc.), the global system budget must be checked and balanced to provide good system performance for all HPC applications. Given constraints on the budget of today's PRACE Tier-0 hosting sites, they are obliged to limit their supercomputer electricity power consumption in the range of 2 to 5 MWatt. One can assume a moderate increase of the power envelope in the future in the order of 3 (6 to 15 MWatt) **[M-ARCH-7]**.

The successive milestones relative to energy efficiency, as specified in the first version of the SRA (100, 45 and 20kW/Pflops), should be reassessed, as technological difficulties arise, they are being delayed. The first 100kW/Pflops milestone has been met at the end of 2016 (2 years delay) as described in the green Top500 list of 2016. The gap between peak performance and real applications performance is growing; as a consequence the exascale performance target cannot be anymore defined based on linpack[86] benchmark execution. It is now defined as 100x more performance for relevant applications compared to today's state-of-the-art PRACE Tier-0 systems. These provide a theoretical maximum throughput of floating-point operations for real applications in the order of 10 PFlop/s (one cent of one Exaflop). The performance metric is, however, not limited to throughput of floating-point operations and might be different depending on the type of application performance improvement must be measured on relevant app **[M-ARCH-8]**.

[86] https://www.top500.org/project/linpack/

6.1.5 Virtualisation

Virtualisation is an important tool for improving HPC systems ease-of-use, reliability and security. At the node level, containers can be set-up to facilitate system administration. Containers will provide a flexible way to tailor the run-time environment for each user and application. They will also enforce better security as applications will be insulated from system software and other applications running on the system. When full insulation and very different OS's are required, hypervisor based virtualisation support remains the solution.

At network level, virtualisation support will allow a better Quality of Service (QoS). It will arbitrate between concurrent users, applications and data flows and their respective priorities. Another important aspect, it could help improve system resiliency with an easier implementation of Check-Point/Restart at the system level.

It is interesting to develop virtualisation at all levels of the HPC systems and in a coherent way.

6.1.6 New application domains

High Performance Data Analysis (HPDA) is a good example of new application domains which could benefit from the HPC experience. HPDA applications have emerged in recent years and have been, so far, mostly based on the Map-Reduce distributed algorithm. New classes of problems are now requiring more sophisticated approaches (e.g. graph analysis and real-time analysis), and the necessary tools, HW, SW and development environment, are being investigated. There seems to be a lot in common with HPC as these new algorithms require a much tighter programming environment. HPDA could make use of HPC technology. And HPC architecture could take into account HPDA specificities.

6.1.7 New disruptive HPC architectures

Besides HPC accelerators (GPUs, Many-cores, etc.), other types of processing elements such as FPGAs, DSPs, etc. have also been proposed for various dedicated applications. They seem well suited for in-flight data processing. However, the specificities of their programming model have limited their adoption so far. The situation must be reassessed in light of the new developments underway, which integrate these devices more tightly with the rest of the system's resources.

New architectural approaches for enabling near-data processing, such as processing-in-memory or in-network-processing, could help to mitigate data transport bottlenecks and reduce data transport. The latter could help to improve energy efficiency as an increasingly larger fraction of the consumed power is spent on data transport.

Another interesting research area that could lead to disruptive changes in system architectures is photonics. Full optical switches supporting new optical networks could completely change the way the system's resources (processors/memory/storage) are organised.

Ultimately, completely new architectures could be proposed for HPC systems in general or for important subsets of HPC applications.

6.2
SYSTEM SOFTWARE AND MANAGEMENT

6.2.1
Operating Systems

The purpose of an operating system is to efficiently manage the resources of a node and to bridge the gap between the actual physical resources provided by a computer system and the runtime system needed to implement a programming model. Given the rapid change in resources and programming models, a basic operating system must be defined for the Exascale community. A common set of APIs will be defined, that could be used by a runtime system to support fully autonomic management of resources.

The node architecture in supercomputers is becoming more and more complex. Increasing levels of parallelism in multi- and many-core chips, complex memory hierarchies and emerging heterogeneity of computational resources, coupled with energy and memory constraints, force a re-evaluation of our approaches towards operating systems and run-time environments. Specifically, new NUMA topologies for memory and I/O, and runtimes that enable scalable and efficient multi-threaded concurrent task execution are worthwhile research and development topics. The innovative block and byte addressable capabilities of NVM technologies, and the deeper memory hierarchies delivered by technologies such as MCDRAM and HBM, have to be properly supported at the OS and driver level, and the option of providing a unified addressing space should be explored to support new data-oriented programming models and optimised PGAS implementations **[M-SYS-OS-1]**. Data-aware kernel scheduling policies and memory management policies for the support of Byte-addressable NVM are other relevant topics in this area.

Use of computer resources for combined HPC and data analytics workloads, or sharing the same infrastructure between these domains requires specific functionality such as virtualisation support, data centric intra-node scheduling, and soft real-time capabilities. Advances in storage and I/O subsystems help here, yet they have to be made visible and usable through system SW and OS extensions.

To make the variety of accelerators that are either available on the market or announced (GPGPUs, FPGAs, machine learning accelerators, etc.) available for applications and their use as easy as possible, specific support for emerging offload programing models in the OS space and in user-space should be evaluated; in addition, specialised containerisation and microkernels can play an important role, and research in this direction is encouraged **[M-SYS-OS-2]**.

Besides the ubiquitous PCI Express standards (with the latest version 4.0 now becoming available in first products), a number of (partly competing) international standardisation activities are underway which target storage and accelerator interfaces both within a node and between nodes. Examples are CCIX, Gen-Z, and OpenCAPI; at this point in time, it is not clear which of these will succeed in being adopted for the majority of relevant platforms, whether different standards will be brought forward, and where the balance between the specific benefits brought by standardisation (interoperability and stability) and the promises of very tight proprietary (on-package or SoC) integration will end up to be. For the OS and driver software this means that potentially a mix of standard or proprietary interfaces will need to be supported.

One of the remaining great challenges for future OS development is a holistic and consistent definition of OS and runtime system interaction with the objective of minimizing energy use, maximizing fault resilience and performance plus scalability. This includes the avoidance of OS jitter. Existing work is looking at parts of the problem only, and this subject as a whole is still unsolved.

Finally, system security is a global critical topic; experience shows that security provisions have to be re-evaluated and sometimes updated with each significant new hardware or SW development. Today, use of SE Linux with appropriate provisions for system monitoring and intrusion detection is the state of the art. Driven by the needs of server computing at large, hardware and system software solutions for strengthening system and data security are emerging. Examples are secure enclaves for storing keys or other private data, encryption of all contents stored in memory, secure partitioning of systems between containers and/or VMs, and advanced authentication, authorisation and accounting mechanisms. As it was the case in the past, HPC systems will take up mature solutions in this space.

Specific requirements for HPC and HPDA arise from the need to protect sensitive data (for instance medical data), the use of special, often user-space communication protocols, and the common practice to have multiple tenants run at the same time on a large Cluster system, each one in a separate Cluster partition. Research into strengthening the security of the key communication layers (such as MPI and PGAS implementations) to bring them up to the same standard as common data centre communication solutions and into isolating system partitions from each other to avoid leakage of data or meta-information will be required. Research into leveraging in-built security functionality in new hardware (such as the automatic encryption 3DXpoint memory promises) in HPC and HPDA job and workflow scheduling, data management, and user-system interactions is also required **[M-SYS-OS-3]**.

6.2.2 Interconnect management

In HPC, interconnect development has been driven by raw performance, in particular by achieving lowest latency and highest bandwidth. The steady improvements of advanced Ethernet implementations will keep the pressure on, with Ethernet bandwidths today rivalling what can be achieved with specialised interconnects such as Infiniband or Omni-Path Architecture, and significant R&D into reducing Ethernet latency showing initial encouraging results.

Besides raw performance, there is a clear need for progress on network virtualisation and QoS both for achieving dependable network performance for each workload and for isolating partitions of a large system from each other.

Dynamic Interconnect topology management and routing is expected to react on message requests from applications or I/O in real-time and create optimal traffic patterns. This requires advances in routing algorithms (such as truly dynamic routing), and for performance reasons needs to run in user space, bypassing the OS. This will lead to a new definition of the borderline between low-level system control in the OS and application use of user space interconnect interfaces. Congestion avoidance and trouble-shooting diagnostics of advanced interconnects will be key to deliver reliable end-to-end performance, satisfying application SLAs and therefore guaranteeing network QoS. The logical conclusion of this is the investigation of the uptake of software-defined networking techniques from the telecommunications and general data centre compute world.

At the interconnect adapter-level, driver and low-level interfaces will evolve depending on hardware technology (in particular closer integration of adapters and nodes) and on new programming model needs. Here, a tight cooperation between system software developers, system architects, and programming model designers will be necessary to integrate new low-level protocol capabilities and define a well-adapted API for higher levels of programming software.

The evolution of storage-class memory and its tighter integration into systems tailored for both HPC and data analytics will lead to new usage modes and functionality of interconnects. Examples are remote access to storage (such as NVMe over fabrics) and memory. RDMA evolution, over fabrics operation, peer to peer communication are great integration challenge for interconnect **[M-SYS-IC-1]**. The support of a virtualisation environment over the entire HPC infrastructure and software defined network capabilities is required but not yet fully integrated.

The emergence of new on-node link technology with very high bandwidth and fabric features such as NVlink creates the need for managing a hierarchy of interconnects, with the required specific investments into federated APIs and tools for performance monitoring, congestion control, dynamic routing, topologies, and diagnostics.

6.2.3 Cluster management software

The evolution of Cluster Management tools to introduce on-the-fly data analysis and post-mortem data mining did start in 2015, yet is not complete. Near real time, event-driven health-checking and introspection is vital to achieve stable operation of large systems. Modularity and heterogeneity of such systems will require the development of a powerful data integration model for system events and performance data. Maintaining a clear view of the configuration and status of HPC systems is indispensable as they are exceedingly complex and susceptible to small perturbations having an extraordinary impact on performance, consistency, and usability.

Today, the ability to manage millions of components is provided by innovative cluster management framework. The new

challenge of cluster management is much more driven by the multiplicity of application execution contexts. Virtualisation and containerisation approaches offer a way to define and enforce specific execution contexts [M-SYS-CL-1]. Cluster management software needs to integrate the management of these virtualised resources, and should achieve interoperability with established Cloud management solutions.

To accommodate the growing importance Cloud computing, new, efficient and secure methods to include HPC resources in private and public Clouds have to be developed. These methods should apply both to traditional HPC applications and to data analytics and AI/ML workloads.

6.2.4 Resource management and job scheduling

Resource management, orchestration and task/job scheduling are critical parts of the software stack for HPC and Big Data use cases alike **[M-SYS-RM-1]**. Improvements in several directions compared to the state of the art will be necessary to reach highest energy efficiency, scalability and system flexibility and throughput across both use cases, and in particular for applications that combine HPC compute elements with large-scale data analytics. Specifically, data has to be treated as a first class resource in determining resource allocations and job/task placements, and additional criteria such as network topologies, guaranteed interconnect performance, optimal CPU architectures/types and kinds of accelerators have to be taken into account **[M-SYS-RM-2]**. Location of data that is required by a task job, either on I/O servers or in NVM of certain nodes, will drive allocation, as will proximity and guaranteed throughput to storage systems for output data. The objective here is to avoid the transfer of data as far as possible, and thus save energy and speed up the time to solution.

Emerging HPC and HPDA usage models do introduce new, dynamic and flexible execution models, and require the flexible and efficient orchestration of complex, multi-step workflows (M-SYS-RM-3). Co-allocation of work where a previous workflow step has created data will become important, as will the sharing of in-memory data between workflow steps.

Orthogonally, to be able to achieve data-aware and power efficient scheduling, applications consumption profiling and resources and data location and affinity must be integrated to scheduling policies. Mechanisms for true dynamic resource allocation, orchestration and scheduling will play a key role for achieving efficient execution of workloads with time-varying resource needs, or of malleable applications that can gracefully react to changing resource availability or workload priorities. This will require the near-real-time collection of fine grained information on job/task execution **[M-SYS-RM-4]**.

To support the growing variety in usage scenarios and required SW environments, efficient resource provisioning is quickly gaining importance. Resetting or rebooting nodes can take an inordinate amount of time, consuming energy and delaying the execution of jobs/tasks. Thus, mechanisms to quickly provision the right VM or container environment should be investigated.

Two recent proposals for handling heterogeneity and supporting workloads with resources that exactly match their requirements should be mentioned here. The concept of a "modular supercomputer" relies on the federation of several sub-Clusters (or modules) with different characteristics (e.g. GP computing vs. throughput/SIMD vs. reconfigurable computing) through a high-performance fabric and on composition of sub-Cluster partitions to match the resource needs of heterogeneous applications **[M-SYS-RM-2]**. On the other hand, the concept of resource disaggregation as proposed by Facebook's Open Compute project and Intel's Rackscale Design Architecture composes a virtual system for each workload out of the basic disaggregated resources (such as CPUs, accelerators, memory/storage and network channels) from the ground up. Both approaches provide opportunities for extensions in the runtime system and resource management areas, in effect creating the equivalent of software-defined infrastructure for HPC and HPDA workloads **[M-SYS-RM-3]**.

Finally, the immense scale of Exascale machines raises the question of whether centralised control and management of their resources is at all feasible. Research into hierarchical and local resource management techniques which address this scalability and reliability problem is encouraged.

6.2.5 Visualisation software

With the increasing amounts of data produced by HPC simulations and large instruments, advanced visualisation techniques become critical to enable scientist and engineers to derive insights from a sea of data, leading to new scientific

discoveries or improved engineering artefacts and products. The same holds true for many large data analytics workloads – the findings of these analysis codes can themselves be very complex, and an intuitive visualisation greatly helps to understand results; in some cases, a skilled data analyst uses advanced visualisation to navigate seas of data, discover structures and then comes up with an automatic analysis approach. 3D Visualisation in itself is quite a mature field, with firm standards and APIs that do save most of the use cases. Extreme scale HPC and HPDA push the boundaries here due to the sheer volume of data (which cannot be held in a single node's memory anymore), and the highly complex and, in the case of HPDA, high dimensionality of data.

Therefore, new approaches and tools should be developed that are able to cope with high dimensional data, very large graphs and other highly complex topologies. For the more traditional visualisation of spatially arranged data, techniques for volume visualisation and fully realistic rendering of 3D structures have to be developed. It is critical to give the end users a full range of visual cues to help instinctively understand the spatial relationships and properties of data. Such visualisation has to be fully interactive, usable in a Cloud infrastructure setting, and has to able to cope with data updates in interactive frame rates. To ensure the 3D realism and provide headroom for adding detail, such tools should use Ray-Tracing methods.

Some of the current applications can no longer store data copies (even for a single timestep) for a visualisation tool to ingest and process. Instead, in-situ data extraction and visualisation is required; expectation is that with the growth in data volumes, this capability will be critical for many more use cases **[M-SYS-VIS-1]**. Advances in non-volatile memory technology could be a good basis for a scalable in-situ visualisation setup, and programmable ways to extract only a relevant subset of the generated data will increase the efficiency of the visualisation, in particular across a remote Cloud connection.

Visualisation of data that is distributed over many thousand nodes poses scalability challenges besides rendering; here, scalable compositing is one of the key issues to be investigated.

6.3 PROGRAMMING ENVIRONMENT

The development of programme ng environments to support the growth of HPC applications and their transition to exascale has a focus on high-productivity for application developers and the efficient and interoperable integration within the complete system architectures of developing and future HPC systems. In addition, the convergence of HPC and HPDA can be supported by programming environments that facilitate the integration of, or interoperability with, proven HPDA methods and frameworks.

Improved productivity for application developers can be addressed by the reduction of programming complexity through advancements throughout the programming model and system software stack. Potential approaches for this include increased intelligence throughout the programming environment and higher level abstractions allowing separation of core algorithmic issues from implementation and optimisation concerns.

In addressing the transition to Exascale, scalability of the realisation of programming models and software tools at all levels is a key concern. The development of programming models and associated runtime systems and compilers should be done in a co-design activity together with highly relevant applications (and, naturally, computing systems), with the aim of ensuring code maintainability and functional (and, wherever possible, performance) portability across existing and possible future architectures and systems. Programming models and frameworks that expose natural algorithmic asynchronicity would contribute to those goals. The need to facilitate efficiency and scalability for legacy applications is of particular importance.

Programming models and their runtime system components should take information about energy, load-balancing, communication, locality and data accesses, significance of the computation, checkpointing needs and similar into account. The required support from the system software level, e.g. through resource managers, for those issues has to be investigated and documented. There should be strong interoperability throughout the programming environment, including compiler tools and runtime systems, debuggers and performance

tools, linking information to the programming model and source code. A plurality of approaches should be pursued with the aim of ensuring that the tools provide insights into the way production codes work on large systems sufficient to identify needs for optimisation for performance or energy use and enable the application programmer to implement these. Furthermore, software tools ensuring program correctness in complex scenarios, such as for asynchronous models, are also in the scope of enhanced handling of program complexity.

There should also be a path towards long-term formal or de-facto standardisation of the programming models and APIs in order to encourage their adoption in industrial and scientific production codes. Indeed, there should be an emphasis on production-grade deployment and sustained support for the models and frameworks developed.

HPDA applications have not previously been covered by the HPC programming environment, one reason being the relatively recent increase in importance (though some of the algorithmic approaches do indeed have a longer history) and very rapidly changing landscape of programming languages, tools and software systems. It is however recognised that a convergence of HPC and HPDA offers great opportunities and that there is a significant potential in the provision of full interoperability, the identification of commonalities in the software stacks and opportunities for significant optimisations, and in the possible provision of "cross-area" programming environment components. The investigation into these possibilities should include a thorough analysis of data analytics tools (such as Hadoop, Spark, etc.) and machine learning frameworks (including deep learning frameworks such as Torch, Caffe, Theano).

6.3.1 Innovative & higher-level parallel programming

This topic covers approaches targeting increased productivity of application development through complexity reduction and includes in particular research into multi-level and domain-specific language frameworks as well as dynamic workflow systems.

The separation of algorithmic vs implementation concerns is a key approach to reduce programming complexity for the application developer. The aim is to provide an abstraction of the underlying computational algorithms/numerics (which are typically hardware-neutral) separated from the actual data structures and parallel programming/runtime system implementation (which must be adapted to a specific target hardware system). Possible approaches to realise this are: meta-programming and scheduling environments building on auto- and self-tuning parallel libraries, and the use of domain-specific language (DSL) frameworks. Code generation can be used to implement decisions regarding the best data structure and implementation choices, in order to take full advantage of a target extreme-scale system. An intrinsic aspect of the high-level programming approach is that specific hardware (and related run-time software) features such as accelerators and near-memory/near-storage processing would be supported in a way that is transparent for the application developer.

A particular need for DSL framework research is the avoidance of a proliferation of DSLs through the identification and use of common core architectural components and common, re-usable features, leading to a small number of more generic DSLs and their integration into more general purpose programming environments.

At the workflow level, there is a need for application-independent dynamic workflow systems that enable the integration of HPC simulation and modelling with data analytics. Such workflows are expect to be composed of HPC simulations, data analytics (at the input, interleaved with computation, or at the output), visualisation and persistent storage/databases. The workflow should be dynamically instantiated, enabling dynamic deployment of new simulations or computations. Such workflows should provide an end-to-end coordination layer that supports streaming inputs and outputs.

The identification of commonalities with HPDA applications and realisation of support in HPC high-level programming systems would be extensions of the goals described above.

Related Milestones are: **M-PROG-1, M-PROG-2, M-PROG-3**

6.3.2 Effective interaction with the runtime system

The effective interaction with the run-time system requires the appropriate APIs to the applications (or high-level application environments discussed above) in order to transfer information (application metadata) between the application and the computing system as well as to be able to realise the

computational schemes that best exploit the system. For the latter, key aspects include data layout, data movement, dynamic load balancing, resiliency and the ability to dynamically adapt to changing resources and application needs, thus including pro-active resilience methods. Using suitable abstractions at the application level, optimisations can be realised by the runtime system. Programming tool intelligence should be based on cost models that propagate (throughout the software stacks) information about energy, load-balancing and communication requirements. The information transfer (from API through to the runtime system) needs to support flexible run-time hierarchies occurring in dynamic and heterogeneous systems.

Related Milestones are: **M-PROG-1, M-PROG-4**

6.3.3 Interoperability, composability and standardisation

In the first instance, Interoperability refers to interoperability between programming models. Interoperability between programming models tackles heterogeneous systems with reliance on specialised APIs and allows the programmers to change only performance-critical parts of legacy codes without having to change the whole communication kernel. Furthermore, interoperability also refers to performance tools, debuggers, verification tools, and run-time systems that understand the programming model abstractions. In addition, it refers to the interface(s) with scripting languages and workflow tools, couplers, and the use of persistent objects.

Composability is the ability to build new programming models out of existing programming model elements, leading to hybrid programming models. Single applications could then combine the use of different programming models to enhance usability and achieved efficiency. The various "components" (including the run-time system) should cooperate among themselves and with the system software to efficiently exploit the shared physical resources.

Both interoperability and composability aspects are particularly important for PGAS and task-level developments, wherein Europe demonstrates strength.

The support of the latest relevant standards is extremely important (including in particular C++17, Fortran 2015, MPI 3.1, OpenMP 4.5, OpenCL 2.2) while there is a need for long-term formal or de-facto standardisation of programming models and APIs in order to encourage their adoption in industrial and scientific production codes. Since interoperability between program models and runtime systems has a broad impact, the standardisation of resource management has a high priority as has the standardisation of the use of memory and storage hierarchies. The growth of HPC applications areas, such as those exemplified by the Centres of Excellence (CoEs), implies that a broader scope of standards should be considered; the need to consider the convergence with HPDA applications will significantly extend that scope.

Related Milestones are: **M-PROG-1, M-PROG-2, M-PROG-3**

6.3.4 Performance Analytics

There is an increasing need for intelligent performance tools (including analysis of energy use) since the goal is to obtain real insights on production codes, where the number of threads and quantity of data makes it impossible to understand a traditional trace. The increase of data to be analysed and to be displayed comes both from the growing complexity and scale of production codes as well as from the number of key parameters (energy, vectorisation rate, bytes/flop from each memory level, etc.) to be considered.

The gap between the tool's output and the necessary source code changes should be reduced, by mapping to the source code structure. For example, current tools allow energy metrics to be displayed, but these are not linked to application codes sections or features. High resolution energy metrics should be used at different levels: at the exploitation level in order to optimise the use of the whole system and from the application/algorithmic point of view in order to optimise the ratio of flops/watt.

The scalability of the tools should be improved by using techniques for data reduction. Beyond that, there is a need for tools to provide predictive capabilities for application scalability. There is a need for holistic tools across HPC and HPDA and support for the tools ability to understand code execution on highly dynamic systems. The use of machine-learning or artificial intelligence approaches could deliver improvements in user experience.

Related milestones are: **M-PROG-6, M-PROG-7**

6.3.5
Debugging & program correctness

Debugger technology is needed which can support applications that have been developed on and for dynamic, heterogeneous computing systems, using both current and non-conventional programming models, languages and APIs, and deployed on the full range of target systems towards exascale; the complexity in debugging applications is growing due the increased use of relaxed consistency models for scalable execution. The interoperability issues discussed earlier also generate requirements for debugging tools and the theme for effective back-reference to the application source-code is similarly important, for example by model-centric debugging.

It should also be noted that not only do application developers need debugging tools, but run-time system developers also need debugger support for the development of scalable software systems.

A key topic with increasing importance (as application complexity grows) is program/software verification. Advances have been made in parallel program verification/verified programming, including models that account for cost measures.

Related milestones are: **M-PROG-6, M-PROG-7**

6.4
ENERGY AND RESILIENCY

Improving energy efficiency by several orders of magnitude is a prerequisite on the path to Exascale. The research which is necessary in the context of 'energy' needs to target four different dimensions: (1) reducing the "innate" electrical power consumption of the next generation HPC system's hardware infrastructure, including the multiple power conversion steps, (2) more efficiently extracting the heat generated, (3) reducing the data centre overall power consumption and (4) make applications more resilient to the faults in the system.

Based on the HPC Challenges we have established the following milestones for the years 2018-2022:

6.4.1
Characterisation of computational advance as function of the energy/power metric

Most common computational metric in HPC is Flop/s. This may not reflect all computations; some application dependent measure is needed to quantify advances per job, subroutine or block of code. A way must be found to measure energy and power spent for the chosen computational metric. Reliable statistics for application dependent measurement must be presented for processor and accelerated computation. This may require development of additional sensors and measurement methods.

The goal of this research is to deliver application specific measure of computation done per unit time and energy spent and includes the relevant data transfer metric as much as network and storage are involved in the computation. This can be done manually for selected applications, the challenge is to extend the methods to standardise the data collection in (semi-) automatic way. The standardisation may proceed in way of an API or a library that can be used in manual or compiler supported way.

The expected result is the full characterisation of application use of resources as function of the power and the energy spent during the computation. The data assembled should allow formulating the minimisation problem for the application.

Related milestone: **M-ENR-MS-1**

6.4.2
Methods to manage the computational advance based on pre-set energy/power metric

This research area covers methods to manage the energy spent in the computer for the most efficient computational advance. Based on the results gained by the previous research area (see **M-ENR-MS-1**), it should propose methods to set energy consumption to be proportional to the computational advance. This may be an extension to the first research topic, however, taken separately, because the minimisation problem must be solved such as to predict the optimised energy and/or power strategy for the most efficient computation with respect to user policy.

The goal of this research is to obtain a prediction strategy and energy consumption management methods based on the application specific computational advance. The expected result is a library and/or a system to set or to optimise the power and/or energy spent for the calculation of a job, subroutine or block of code in the most efficient way with respect to user policy.

The expected result is a (hardware and software) system that adopts its resources to the application load achieving *Proportional Computing Based* on the selected policy. This proportionality should apply to the computational resources and also to the network and storage bandwidth.

Related milestone: **M-ENR-MS-2**

6.4.3
Throughput efficiency increase by scheduling cores/ functional units within processor

Methods to manage the growing number of cores and functional units within processor as a way to exploit the limitations of the processor power envelope (the "dark silicon"). This may be done from OS/Compiler site or from within the application when processor vendors start offering the relevant tools. The dynamic load balancing may exploit idle time to send instructions to functional units and delaying the non-critical instructions that do not fit the power envelope.

The idle time is an inevitable consequence of resource constraints; the minimisation of idle time is known as the "Slack Allocation Problem" and is widely researched. It is expected that with the growing number of the cores/functional-units the opportunities for optimisation will also grow and the increase of the efficiency of computation should be targeted by this work. This research topic asks for methods and procedures to manage computational resources and/or balance the usage of the computational resources using energy and/or power in the decision tree.

The goal of this research topic is to demonstrate increased throughput of computer by execution of additional code using idle time of programs or balance computation such as better exploit the available resources. Implementation of this paradigm will increase the overall efficiency of the computer system, as more jobs will be able to pass per unit of time within the same power constraints. Research in this area may be helped by the development of work scheduling methods, (see **M-SYS-RM-3, M-SYS-RM-4**).

The expected result is a description of methods and practical demonstration of a system to run additional code without slowing down execution of other programs.

Related milestone: **M-ENR-MS-3**

6.4.4
Optimisation of the energy spend by the facility by controlling the coolant temperature down to the device level and taking the infrastructure energy cost into account

We expect innovative architectures to appear with new levels of memory hierarchy, the 2.5D and 3D processor and memory integration, optical networks on or close to the chip package. This will bring in the new challenges in cooling. This milestone should establish a procedure to control the cooling to minimise heat dissipation at the device level. For example, it may establish practical way to use the 2-phase liquid cooling for a single device, physical nodes and beyond.

The goal of this research is practical demonstration of improved heat extraction from complex chip packages coupled with techniques to regulate the temperature of the coolant to optimise the overall power and energy efficiency from single device to the whole installation with many of these devices. The minimisation should include the infrastructure cost, such as the cost of refrigeration (if used) and the pumping cost.

The expected result is practical demonstration of working installation with management software that can optimise the temperature to cool individual chips while minimizing the overall energy cost.

Related milestone: **M-ENR-HR-4**

6.4.5 Collection and Analysis of data from sensor networks — the Big Data challenge

Collection and Analysis of statistics related to events and measurements in computers. Sensors provide a fast data stream that must be stored and processed to extract valuable information on the function of the facility.

A fast database of log statistics exists, but these are not immediately related to failures. Classification of failures is needed and events prior to these failures need to be explored for patterns that may correlate to failures. Obvious data to look for are the temperatures and voltages on the hardware boards, but other measurements may be needed for better prediction strategies.

Such work has been done specifically for memories and disks, where failures can be easily detected. A better and more complete classification is needed to account for all failures, including software errors that may be correlated to hardware events in the system. Statistics on the silent data corruption is also very sparse. Due to unpredictable and rare occasions of some failure patterns, this may be a multi-year effort. However, building a comprehensive sensor network will collect statistics on all events in the machine, including the events related to the infrastructure.

The goal of this research is to collect and manage data from the facility sensor network and devise ways to analyse this data efficiently to discover patterns and trends that may be correlated to fault conditions and discovery/classification of the fault conditions in the system.

The expected result is lists of patterns that probably lead to failures and methods to collect the related measurement and statistics proving the case.

Related milestone: **M-ENR-FT-5**

6.4.6 Prediction of failures and fault prediction algorithms

Prediction of failures in computers and fault prediction algorithms: This is an extension of the research topic **[M-ENR-FT-5]** but is treated separately, as it requires a different methodology. This is mostly a computational/mathematical work needed to develop prediction algorithms based on statistically significant patterns for failures. Such work has been done before, but in absence of statistically significant patterns, the impact is low.

The goal of this research is development of prediction algorithms. The deduction of hazardous pattern may require methods of High-Performance Data Analytics with usage of Machine Learning techniques or other methods to be discovered to predict the possible failures.

The expected result is a system that predicts probability of a specific failure occurrence within a given time frame, if certain events or patterns of events are measured in the system. These patterns will emerge from system software providing feedback on the hardware based resiliency as mentioned earlier. The research should lead to standardisation proposals for the API to manage this process as well as an API to provide the link to the Application Based Fault Tolerance (ABFT) capabilities.

The prediction mechanism may be implemented as system daemon with notification capability and/or as a dedicated hardware device that connects to the facility sensor network.

Related milestone: **M-ENR-FT-6**

6.4.7 Application recovery from fault conditions in the system

The user assisted checkpoint at synchronisation and application restart from the last checkpoint in case of a crash is an established recovery procedure. To increase scalability it is necessary to devise a scheme where only the crashed threads are restarted, or some other ways to avoid the restart of the whole application. Also, the application can be recovered with different resources, as the crashed elements may not be available.

It may be necessary to consider fault-tolerant communication libraries and/or algorithms that can recover from lost resources or lost precision to continue execution. The non-volatile memory may be used to store the checkpoint for the computation and/or there are even more powerful ways to use architectural advances to increase the resiliency of applications.

It may be necessary to invent methods to obtain information when to record a checkpoint, if the checkpoint scheme is used. The information may come from the Failure Prediction (as explored in milestone **[M-ENR-FT-6]**, from the OS that can communicate hardware events through the appropriate API, or from the sensor data, possibly after suitable analysis **[M-ENR-FT-5]**.

If the ABFT recovery scheme is used, this milestone should demonstrate robustness of this approach also on machines where the strong reliability constraint is relaxed due to the subthreshold voltage settings and may require data encoding to detect/recover from the silent data corruption. This should be considered in the framework of new application development (see also M-ALG-1 and M-ALG-2). The expected result is a system that demonstrates survival of application as resources it uses become unavailable.

Related milestone: **M-ENR-FT-7**

6.4.8 Energy/Power efficient numerical libraries

An obvious extension of the research area **[M-ENR-MS-1]** and **[M-ENR-MS-2]** is the use of these techniques for the numerical algorithms or the scientific libraries. Optimisation may be done for either the power efficiency or energy efficiency or both as with the Energy-Delay Product metrics (EDP).

One of the requirements of this research is bit-reproducibility of the results under various run conditions. It is expected that either energy or accuracy can be specified as termination criterion for the algorithm. The full bit-reproducibility is a hard problem. However, less accuracy than 64-bit may be specified, in which case it is easier to obtain. This may be explored in algorithms that increase the accuracy iteratively, or in other ways, trade accuracy against energy usage.

The goal of this research is to advance towards highly scalable numerical algorithms and scientific libraries that can deliver results of specified accuracy with quantifiable and efficient energy usage. A welcome side effect is the tolerance to precision loss coming from faults in the system. Much of this research may rely on the results obtained from algorithms development, such as **[M-ALG-8]**.

Expected results are libraries for linear algebra, FFT and others, demonstrating the desired characteristic.

Related milestone: **M-ENR-AR-8**

6.4.9 Highly efficient HPC installation

Demonstration of a sizable HPC installation with efficiently cooled components. The energy losses must be small and under control. This is a wish list for the work carried out in this research for at least a one-year period of running such machinery:

- Demonstration of the Energy usage minimisation on the machine and energy reuse
- Measurement and demonstration of the common metric such as Power Usage Effectiveness (PUE), Effectiveness of computational energy (TUE)
- Demonstration of the network bandwidth proportionality as well as the optimisation of the storage access pattern
- Comprehensive sensor network with collection of data and analysis of this data feeding to the DCIM software that effectively minimises the energy spend by the facility
- Collection of energy usage data for job statistics and energy saving measures for specific jobs
- Methods to reuse heat produced in such installation and/or using sustainable technologies
- Study this usage-model for Total Cost of Ownership (TCO) and exploitation characteristics of such installation, taking infrastructure into account
- Work related to other Chapters demonstrating progress towards the Exascale systems

· Work related to Research Areas that have not been resolved at the time of the installation

This computer installation should demonstrate PUE~1.05 (i.e. 5% infrastructure load) and TUE~1.1 (i.e. <5% power conversion losses) which would mark a significant improvement in the power/voltage conversion efficiency. This must be entirely liquid cooled facility.

This should take into account the possibilities to optimise the overall efficiency by reducing the coolant temperature to limit the leakage current. This milestone should account for the software to manage the facility for the energy and power capping strategies.

Optimisation of the energy spent to transfer data between the compute nodes (interconnect bandwidth), optimisation of the storage access and LAN usage optimisation is part of the approach towards this milestone.

Expected results are measurements and reports related to energy efficiency and TCO.

Related milestone: **M-ENR-MS-9**

6.5 BALANCE COMPUTE, I/O AND STORAGE PERFORMANCE

As highlighted in the general trends, the key focus for balanced compute & I/O is to continue to address the performance gap between compute subsystems and I/O. The gap since the last SRA is only increasing with even higher thread counts and heterogeneity driven by new processor architectures and newer GPGPUs. For Exascale simulations, we envision billions of threads all simultaneously accessing the storage and I/O subsystem. Apart from simulation outputs, the data driving scientific insights will also continue to be fed from external data sets which need to be pre-processed, analysed and post-processed. Simulations and data analysis will be part of one big workflow. The technical research priorities outlined below focus on these critical requirements and are aimed to provide guidelines for new R&D initiatives in Balanced compute and I/O research embarked upon in the European HPC community. The next goal post for Storage and I/O is indeed getting to Exascale readiness in the 2020/2022 timeframe.

6.5.1 Non-Volatile Memories in I/O Stack & Deep I/O Hierarchies

Traditionally storage for High Performance Computing consisted of just a single tier of storage, with high performance "scratch" disk drives managed by parallel file systems such as Lustre, GPFS, etc. Such systems are fundamentally bottlenecked by the inability of disk drive data accesses to keep up with the I/O bandwidth needs which at Exascale will be exceed hundreds of Terabytes per second. To deal with the inability of disk drives to keep up with the bandwidth needs, the crude solution has been to add more spindles/drives in parallel. However, doing that will not solve the metadata bottlenecks associated with data accesses nor the latency requirements of applications.

Having a Flash tier between the disk drive tier and compute is the obvious solution, which is also supported from a cost and economics standpoint. Such burst buffer based technology is now heavily supported by the large HPC sites and vendor

roadmaps. However, the Flash tier alone offers a very limited set of data persistence and performance points, as Flash lifetime aspects have been continuously researched with some good progress. Deeper I/O hierarchies which consists of Flash as just one of the tiers, but which includes additional tiers such as Non-Volatile Memories should continue to be addressed. This enables the storage and I/O subsystem to address a much wider repertoire of application performance requirements and data retention requirements.

The current constant and rapid pace of hardware innovations leads to software solution lacking maturity, in the sense that fine understanding of the hardware substrate is not yet fully reached. For instance, the aging process of flash based system is still to be handled correctly at the middleware level. Such understanding is mandatory in order to develop relevant predictive maintenance schemes.

Since the last SRA update there are highly limited, albeit viable choices available for Non-Volatile Memories. Research in the usage of these NVMs should continue very actively – supported by the industry push for providing earlier availability of these technologies within Europe. There is also need for the low-level system software to evolve to make use of these non-volatile memories in the I/O stack. Existing applications should be able to make use of these devices without significant modification.

It is still unclear how all the pieces of the deep I/O hierarchies will fit together for exploitation by both existing and new applications. This is an area that should continue to have significant research focus.

There is now trends that indicate that Flash components (and NVRAM in the near future) will be increasingly available within the compute node. System software and storage infrastructure software needs to be adapted to take this trend into account as software now needs to dynamically expose persistent storage pools within the compute subsystem **[M-BIO-1]**.

Larger scale storage systems, for example implemented as deep I/O hierarchies, assume extremely large name spaces. Such volumes of name-space data could potentially follow a similar trend as other data. As an example, it is possible that for an Exascale system, name spaces could be hierarchically organised with non-uniform access times.

Hierarchical storage could indeed be reconsidered as non-uniform storage access. Due to the diversity of applications, access patterns and requirements, both in terms of performance and of functionalities, the hierarchy of tiers might evolve towards more complex distributed interconnected components with each component presenting a different functional and performance footprint. The front-end of the storage hierarchy will have to drive both data plane and control panel through the interconnected graph of these storage components. Those components could either be software defined or hardware systems. As a final point, backward compatibility in terms of exploiting these deeper hierarchies including non-volatile memories needs to be looked at for existing HPC file systems and storage software infrastructures. Adding in new levels within the storage hierarchy does not mean that existing ones will disappear. It remains extremely important that the community pursue its efforts in deploying new storage and I/O software well integrated with existing parallel file system solutions.

6.5.2 Data Centric Computing

Traditionally storage and I/O infrastructures for HPC have been separate from the infrastructure for big data analytics. However there is continuing trend for the requirement of a common infrastructure that caters to both extreme compute as well as big data analytics type workloads. This convergence is evidenced by big science workflows, which consist of external data ingest, data pre-processing, simulation input, simulation output, data post processing and data archival steps. The I/O subsystem should not just deal with checkpoints from applications but also deal with pre/post processing and data analytics for solving a given problem. Data takes centre stage in such a "data centric computing" infrastructure. There is potential for insights generated from data analysis to be fed back into running simulations. Initiatives looking at a converged infrastructure has picked up more momentum in the last couple of years since we highlighted that first in SRA2. Research should continue to look at addressing such data centric workloads. There are clearly elements of the Big Data analytics software stack that is applicable for data centric HPC and there are many functional overlaps. Opportunities for consolidation exist now more than ever. A clear example is data analytics software frameworks (Apache Flink, Spark, etc.) on top of storage and I/O middleware (such as file system and Object stores) **[M-BIO-4]**.

New architectural implications of data centric computing, for instance, with compute nodes connected to highly distributed mesh of data silos each having their own performance, persistence and reliability characteristics needs to be looked at.

Under the context of data centric computing, a holistic view of energy consumption at a system level should be taken into consideration for smarter decision making on process and function offloading to the storage system - which clearly is expected to provide dramatic savings in terms of data movement energy costs.

In SRA2 we highlighted in-storage compute as an emerging trend. The advent of data centric computing has made in-storage compute/active storage even more relevant since processes, tasks and computational kernels can potentially be offloaded to storage to increase the time to solution.

More critically energy is one of the biggest problems getting to higher scales as we have discussed earlier. Data centric computation will involve a lot of data movement with existing architectural assumptions. Data needs to be pre-staged, post staged, and moved back and forth between the compute and the I/O subsystem. Availability of computational capability within the storage subsystem will alleviate those problems. The software infrastructure for doing pre/post processing (for example: with runtimes) and function offloading needs to be further built out **[M-BIO-5]**.

6.5.3 Object Storage & Cloud Convergence

Object Storage

Object storage software has found its way into HPC having started out in the Cloud. As we highlighted in SRA2 Object storage software exposes a flat namespace and overcomes some of the limitations of POSIX based hierarchical namespaces. The horizontal scaling capability of Object stores is virtually limitless and Object store also provides the semantics to overlay very rich metadata associated with Objects by leveraging infrastructure such as Key Value stores which makes it very easy to search and index data. Given that parallel file systems based data organisation is already reaching fundamental limitations, it is imperative to further refine and adopt Object stores (and associated non-POSIX storage infrastructure software) within extreme scale HPC.

Storage for HPC in the Cloud

With HPC moving to the cloud, there is a need to optimise storage and I/O access in virtualised environments considering that HPC cloud infrastructures will avail virtualised resources through legacy [VMWare, OpenStack, Dockers] and emerging virtualisation technologies for processing I/O and storage. Cloud Object based storage infrastructures are well positioned to provide for the requirements for HPC in the cloud (scalability, multi-tenancy, storage for unstructured data formats, flat addressability, etc.)

Also with data centric workflows creating data in different geographies that needs to be federated, backed up and consolidated (and included as part of HPC simulations performed at the “edge” close to the data sources), technologies to address storage and data management over wide geographies needs to be addressed.

With data sets dispersed over wider geographies and spread across data centres, with infrastructure potentially shared by disparate entities, it is imperative to address data security aspects including encryption and access controls. This issue has seldom received focus in extreme-scale HPC.

Data Security and Confidentiality

Data centric/cloud converged storage architectures lead to data infrastructures accessed and shared by multiple tenants. Furthermore, data has different levels of importance and confidentiality; some can be recomputed, some can be re-generated by recovery schemes while some should not be read without correct permission. It is mandatory in such a scenario to have schemes ensuring end-to-end data protection and security. Such schemes should not be limited just for the data production stage, but has to encompass the entire data life cycle. Even though these data security and confidentiality issues are common across all forms of HPC deployments, they become even more critical for cloud and virtualised environments.

Deployment

Finally, there is a need for convergence between traditional batch schedulers used in HPC for deploying applications and cloud application deployment frameworks.

6.5.4 Feature Rich APIs

There is a need for lower level feature rich APIs that provide different views of data (as dictated by data formats) and enables placement and lay outing of data in different tiers and pools as needed by use cases. These APIs should support a diverse range of applications and their data access requirements considering that there clearly is not a "one size fits all" solution. It should be possible for third parties to easily build additional features on top of these APIs. The APIs should also provide for all the needed features of data centric computing, for example, in-storage compute and function offloading as well as capabilities to extract operation logs, telemetry information, etc. as needed by data management tools and use cases. This is something that continues to be of importance as we progress through the SRAs.

The APIs need to provide portability of functionality across various storage and I/O platforms. It should also be possible for a wide range of use cases to exploit the features of these APIs without incurring performance penalties ("Performance portability").

It is necessary for the APIs to have a clear path for longer term standardisation **[M-BIO-3]**.

6.5.5 QoS & Understandability

QoS

I/O and storage systems should be able to provide guaranteed end-end quality of service for key I/O parameters such as throughput, latency, jitter, etc. Fine grain QoS has still not been achieved even though it is imperative to have this considering that storage and I/O are going to be increasingly shared resources (more so for HPC in the cloud and any associated usage/pricing models, etc). Applications should be able to specify serviceability parameters with Service Level Agreements and negotiate these with the storage and I/O subsystem. Storage and I/O systems providing fine grain QoS should have the requisite characteristics of "just-in-time" self-healing and the ability to continuously track performance with rich telemetry data provided by the storage and I/O subsystem. Despite of having this requirement since the time of the first SRA the community has not progressed much.

There is good opportunity to implement QoS controls in I/O middleware software as middleware implementations (such as parallel file systems and Object stores) typically have wide visibility of storage and I/O infrastructure resources **[M-BIO-6]**.

A very closely related aspect of providing QoS is provision of I/O system resilience and fault tolerance, which we discuss in this section.

I/O System Resilience & Fault tolerance

There are three aspects to storage system resilience. The first aspect is the ability to deal with application crashes, which is typically done by having a good handle on checkpointing methods and rates. The second aspect is data corruption within the storage hierarchy that needs to be handled with error detection/correction codes and redundancy techniques. The third aspect is infrastructure failures (disk drives, networking fabric, embedded processors, etc.)

With regards to all these aspects, we continue to highlight that failures will be a norm rather than an exception at extreme scales. It is highly imperative for the storage and I/O subsystem to deal with constantly occurring faults and failures and be able to dynamically adapt based on a holistic knowledge of the state of the infrastructure. The applications should not perceive any down time because of these infrastructure failures. Such mechanisms need to be built into high availability (HA) systems working for these environments. Further the applications should be able to revert back to any previous known stable states. (Transaction Management). It is indeed important to note that new techniques such as in-storage compute, dealing with new workloads such as deep learning, etc. will put more pressure on the storage hierarchy than ever before.

Understandability: Storage Performance Prediction and Analytics

Different workloads view different capabilities from the storage and I/O subsystem. It is extremely important to characterise the workloads and have the storage system be able to predict the I/O performance for that specific workload. This predictive capability can then be used to study "what if" scenarios through simulation and modelling to study if the I/O subsystem could perform better for the workload under other possible infrastructure configurations and system parameters. These leanings can then be fed back to the real storage system through a policy engine.

To achieve such storage adaptation, it is also crucial to have continuous feedback from the storage system in terms of how all the different layers in the I/O stack are performing at finer levels of granularity for the different workloads. This is achieved through storage system telemetry data that could be continuously obtained through some well-defined interfaces that do not sufficiently exist today as system administrators still resort to unstructured logs and rely heavily on third party tools that do not obtain the necessary system infrastructure telemetry data from the storage system.

Many of these issues can be addressed through a simulation framework (an extreme scale storage and I/O simulation framework), but for which efforts are lacking in the community to develop them. This continues to be a technical priority through all the SRAs **[M-BIO-2]**.

Advanced I/O Benchmarks

There is a pressing need to provide advanced I/O and storage based benchmarks that target metrics such as time/cost & energy to solution for data centric workloads apart from benchmarks that test point parameters such as throughput and latency. It is important for the HPC community to agree on the benchmarks that help provide deeper understandability of storage system responses to a wide class of workloads. Initiatives such as IO500 are some good examples of vehicles that might enable such co-ordination.

6.5.6 Data Management

Data management aspects mainly information lifecycle management mechanisms that includes HSM (Hierarchical Storage Management), Data Integrity checking schemes and tracking of data provenance continues to be a technical priority at this time. HSM mechanisms have only started to be adapted from simple parallel file system-to-archive interfacing to multiple tiers. Flexible mechanisms to carefully track data as it progresses through the different pieces of the workflow (with possible processing steps) and through the different storage tiers/infrastructure in the realm of extreme scale HPC is still missing **[M-BIO-7]**.

6.5.7 Adaptation for New Programming paradigms & Workloads

The storage and I/O subsystem needs better co-design with data centric use cases which include, HPDA and Machine Learning use cases which are emerging as the new class of workloads in HPC. For instance, the needs for storage and I/O for deep learning/AI are not clearly understood even as they are increasingly using HPC resources at the time of drafting this SRA. Non-volatile memories will play a big part for these workloads, under memory constraints, especially in storing input data, weight parameters and activations required for neural networks. It is important to get good examples of these new workloads for meaningful co-design - either within the existing HPC ecosystem or more so, in collaboration with large scale cloud providers who tend to have better examples of such workloads.

Further, very strong interlock is needed with the Centres of Excellence (CoEs) set up during the EINFRA-5-2015 program, which are primarily driven by the application users and owners, covering the full scientific and industrial workflow. Continued interlock is also needed with the new CoE pilots not currently covered by the existing program **[M-BIO-8]**.

Adaptation of new programming models such as PGAS and MPI one-sided communication to appropriately make use of components in the storage stack (especially non-volatile memories) will be very useful for the extension of process address space, especially under memory constrained environments. Introducing such features directly in programming models bypassing middleware I/O layers that introduce overheads will be a very interesting area for consideration.

Unified addressing deals with providing memory semantic addressing (Byte or word oriented as opposed to block based) of storage and I/O subsystems. Trends such as memory centric computing in the last couple of years has called into question whether it was possible and useful to access persistent storage as part of extended memory address space. Unified addressing is expected to provide flexibility for applications and help bypass some of the unwanted I/O middleware layers, as the programming models such as MPI can directly reference data in the address space. With the proliferation of NVRAMs especially in the compute nodes, this is expected to provide the flexibility for applications to access NVRAM either as memory or as part of persistent storage. This is a new trend we highlight in this version of the SRA.

6.5.8
Adaptations for new hardware: Processing environments & Interconnection Networks

Processing environments

The storage and I/O subsystem needs to be optimised for new compute environments – especially very low power processing environments and heterogeneous environments such as new generations of GPGPUs and many core processors. This can be done by identifying key instructions for improving performance and reducing energy, within the instruction set, that can be leveraged by the storage and I/O stack.

General purpose processors have been successful in integrating performance differentiators in their instruction set. The storage community should inspect their I/O stack to ensure that new processing environments include these critical optimisations for data access.

Interconnection Networks

More research is needed to study fundamental limitations of incumbent interconnection network technologies for extreme scale/data centric HPC and to incorporate new interconnection networks with provides features such as Quality of Service and the ability to sustain very high I/O performance at scale. There is a need to look at appropriate networking topologies and advanced dynamic routing algorithms as appropriate for storage I/O workloads.

6.6
BIG DATA AND HPC USAGE MODELS

The convergence of HPC and Big Data is a trend that is happening fast and has already started to influence the HPC world. Data is becoming central even for traditional HPC domains in one way or another, whilst new HPC clients are by default data centric. These trends can be seen as a challenge; however, it is much more useful to be considered as opportunities. The HPC community needs to formulate a plan as to how to address them and benefit from them the most.

6.6.1
Performance Metrics

It is clear that the huge data volumes are here to stay and it is not a transient trend, either because the increased fidelity of simulations is creating more data or because Big Data systems will keep "pouring" data into HPC systems. Thus, we need to rethink the optimality criteria which we have been using to design HPC systems. Firstly, data centric applications are typically less computationally intense. Thus, high flop/s or flops/watt metrics need to be completely redesigned. The HPC community has reacted already by introducing the HPCG and the Graph500 benchmark. It is clear though that these need to be enhanced and augmented for the data centric applications, adopting, for instance, benchmarks from Machine learning such as deep learning, support vector machines or similar devices. Secondly, the overall computation can be split in phases that alternate between Big Data and HPC systems, or follow a certain workflow between them. Thus, the overall distributed nature of the computations and the data handling needs to be accounted for in the new metrics.

Related milestones are: **M-BDUM-METRICS-(1-3)** and **M-ENR-MS-1**

6.6.2
Data Centric Memory Hierarchies/Architectures

Moving data is clearly the primary problem. This includes introducing data in the system, but also moving data across the memory hierarchies. It is clear that disk needs to be used only

as a last resort and thus non-volatile memory will become of central importance. In addition, the data structures used in the data analytics algorithms will often be completely different from those used in the compute bound part. Therefore, efficient data structure transformation will be a key factor. In particular, when considering heterogeneous systems, we will need to adopt and further develop solutions that allow data coherency between the various compute engines. The HPC community needs to offer solutions that are easy to be used by the millions of Big Data users and developers for high-performance data analysis (HPDA). Tools for HPDA need to extent common Big Data tools in levering the efficiency and performance of the underlying HPC infrastructure by maintaining generality for complex workflow patterns. Furthermore, the memory hierarchies on the HPC system need to be efficiently linked with those of distributed Big Data systems. Determining efficient data flows that keep as much of the computation local will be a key factor. Thus, the H/W architectures of HPC systems need to reflect all these needs. For instance, local (to the HPC node) non-volatile storage subsystems, coupled with more traditional parallel file systems need to be considered.

Related milestones are: **M-BDUM-MEM-(1-2)**

6.6.3 Research in Algorithms

Research in algorithms that trade computation with data accesses will be of key importance. This means that we need to develop algorithms that minimise data access by relying on more computation/ post processing that remains in-situ in order to reduce data movement and by consequence save energy. To this end, a systematic analysis of key Big Data applications, including Data and Compute intensive ones is required and a breakdown of the data flows to reusable pieces that form a pipeline. Similar activities have led to great advances in traditional HPC applications (e.g. Berkeley Dwarfs). Generally low complexity, communication avoiding and scalable algorithms that discover global properties on data are required. In addition, we need to rethink distributed computation in Big Data applications in view of the availability of HPC systems that are coupled to these. Even theoretical advances pioneered years ago need to be revisited in view of extreme parallelisation potential available today. As an example, we can consider algorithmic parallelisation for Deep Learning, in which the model itself is to be distributed (as well as the training data).
Related milestones are: **M-BDUM-ALGS-(1-2)**

6.6.4 Programming Models

Programming models and languages for data centric computing will also need to become central in HPC. That is to say, programming models and languages that are heavily used in main stream Big Data research and practice (i.e. Hadoop, Scala, Java etc.) can serve as an inspiration for the needed adaptations of HPC practices in view of Big Data. The HPC community needs to offer solutions that are easy-to-use by the millions of Big Data users and developers. Furthermore, mixed programming models will be crucial in bridging Big Data and HPC environments. In terms of enhancing collaboration, programming and deployment models that allow cross-institution collaboration may have much to offer.

Related milestones are: **M-BDUM-PROG-(1-3)**

6.6.5 Virtualisation of HPC

Convergence of HPC and Cloud computing is a crucial step. We have already started to increasingly see HPC users wanting to use an elastic way of storing data and using resources, since this is far more economic and a reality in mainstream data analytics. The question is now how to bind and merge HPC resources with Cloud architectures. Moreover, the increasing complexity of heterogeneous node architectures of HPC systems (e.g. CPUS, GPUS, FPGAs, DSPs, NVME, etc.) inevitable means that some form of virtualisation is needed in order for users to navigate through these complex environments. Thus, research in high performance containers has very important potential.

Related milestones are:
M-BDUM-VIRT-(1-2), **M-BDUM-DIFFUSIVE-1**

6.6.6 Diffusive Supercomputing: Bridging computation and data acquisition

HPC needs to come closer to data, both in its generation and its consumption. Thus, we need to research as to how to develop heterogeneous HPC data processing systems (HPC and Big Data hybrids), which are flexible in allowing breakthrough simulations and data analytics at the same time. This means

that HPC design needs to become much more holistic, considering simultaneously computation, data storage and computation on storage, data acquisition and data serving. We can envisage HPC systems to be used closer to where data is generated and use these 'Big Data' systems as a crucial pre-processing stage of the data analysis pipeline. Thus, HPC system design needs to take into account IoT design requirements. A key example, particularly important for Europe, deals with IoT for the automobile industry. Vast amounts of data are expected to flow from vehicles. Local computation is of course key (computing-on-the-edge), however, a holistic view of the data is deemed absolutely needed in order to fully materialise the benefits. HPC will play a crucial role for the realisation of the full potential at the central role.

Related milestones are: **M-BDUM-DIFFUSIVE-(1-2)**

The aforementioned trends and target areas, which clearly show that the new HPC uses which are data centric, will demand research on the full hierarchy: algorithms, system S/W and H/W, languages and programming models and storage. Thus, we see this subject as a horizontal one, running across almost all other areas of research focus rather than an isolated vertical theme.

6.7 MATHEMATICS AND ALGORITHMS FOR EXTREME SCALE HPC SYSTEMS

The development of future HPC architectures is strongly driven by several technological trends which in particular adhere to power constraints. New mathematical methods and algorithms are important ingredients in ensuring efficient usability of these future architectures and technologies, as well as scalability from the mathematical methods level, through algorithms down to system levels. They have to cope with increasing parallelism at the growing number of levels, as well as with a deepening of memory hierarchies that can also be heterogeneous. As clock speeds may for reasons of power efficiency be as low as 1 GHz, a computer capable of performing 1 EFlop/s peak performance will have to execute 1 billion floating-point operations concurrently. This would be a significant increase compared to the Sunway TaihuLight system in China, which is at the time of release of this SRA the fastest supercomputer according to the Top500 metric. This system sustains almost 93 million floating-point operations per clock cycle using more than 10.5 million cores. Parallelism is expected to increase at many different levels, with most added parallelism expected at a node level. With power consumption becoming a major factor in total cost of ownership at almost any location, also unconventional hardware architectures, based, e.g. on FPGA designs or near-data processing approaches, e.g. in-memory or in-network processing, that allow for extremely energy efficient, and possibly application specific acceleration, are becoming a competitive option in important application areas. Once again these solutions rely on concurrency becoming more abundant, and thus, there is a need for pushing scalability as well as efficiency limits of applications, with an emphasis on industrial and commercial organisations in Europe, to improve their products or their usage. This is particularly challenging for those cases which are not embarrassingly parallel and feature a high level of irregularity. To keep the system architecture balanced in terms of compute versus, both, memory performance and capacity, a deepening of memory hierarchies is to be expected.

Algorithms and mathematical methods will have to play a critical role to enable efficient exploitation of these new architectures. The practical challenges in algorithm development

are as a consequence changing and becoming more diverse. Beyond optimisation for time-to-solution, optimisation targets like reduced energy-to-solution have become more important.

A vital area for research and development is where the need for scalable computing resources is linked to huge amounts of data, which are generated through simulations or originate from external sources, e.g. data and compute intensive problems, and also, where the need for new mathematical methods and algorithms for extreme data challenges is critical.

Sustainability, robustness and ease-of-use are further aspects which need to be addressed. The sustainability issue is related to the use of standard programming enabling long-term portability of both performance and functionalities across multiple generations of HPC systems. The growing importance of computational methods in industrial processes leads to a growing importance of robustness and reliability. Robustness has multiple facets including numerical robustness, propagation of uncertainties in complex work-flows, or robustness with respect to undetected hardware and soft errors. For broader uptake of HPC solutions, ease-of-use is a critical aspect and should facilitate uptake of new algorithms and mathematical methods by Software-as-a-Service (SaaS) providers.

6.7.1 Robust methods and algorithms enabling extreme scalability

To exploit the performance of future massively-parallel architectures, for many problems new algorithms are required which allow for: a fixed problem size, the increase in the level of concurrency and facilitate the usage of different levels of parallelism or specific hardware accelerators. New mathematical methods may lead to innovative approaches in the use of computation in problem solving which generate new levels of concurrency. Algorithms may have to be hierarchical to reduce communication as well as synchronisation and to simplify dynamic task scheduling. On the other hand, performance variability of the system components is expected on future exascale systems. This makes global communication hiding algorithms such as pipelined Krylov solvers, even more important. At some point a bulk synchronous approach may not be a feasible option anymore. Run-time instantaneous response to the system variability will be required, thus, new dynamic parallelisation and load balancing capabilities should be included on the algorithms design. Mathematical methods such as stochastic and hybrid ones (stochastic/deterministic and deterministic/deterministic ones), including hybrid Monte Carlo and quasi-Monte Carlo, and, asynchronous methods and algorithms, as well as multilevel multiscale hybrid Monte Carlo, domain decomposition methods and algebraic multi-level/multi-grid methods, and mathematical developments that reduce dimensionality (e.g. tensor-train decomposition) are of particular interest. They may lead to innovative approaches in terms of increased scalability of the methods and algorithms enabling reduced communications. They also may result in efficient algorithmic resilience and fault-tolerance. The latter renders them very suitable in problem solving for large class of problems that generate new levels of concurrency. The challenges of emerging exascale architecture can be met by enabling runs with mixed and variable precision, maintaining a high level of parallelism and through multi-level and multi-scale approaches, thereby, matching the hierarchical topology of new computer architectures, which may include compute nodes comprising accelerators that are both, highly parallel and highly hierarchical.[87] These efforts are to be also considered in the context of multi-physics workflows and advanced strategies for coupling several scalable applications.

Finally, an important aspect to be addressed is robustness. Robustness has several aspects including but not limited to the following:

- Fault resilient algorithms which enable the algorithms to run on large and error-prone systems, reproducibility of numerical results and numerical stability of algorithms.

- Additionally, robustness enables many opportunities to trade performance for accuracy, which makes it possible to consider approximate computing scenarios where the exactitude of some computations can be tuned to maximise compute or memory access performance, whilst keeping the results' quality within an acceptable margin.

- Assessment of robustness and fault resilience implies the definition of appropriate methodologies for faults injection and changes in precision, using a representative set of kernels for the exhaustive testing.

Here, also data-related uncertainties need to be considered and addressed, e.g. within a Verification Validation and Uncertainty Quantification (VVUQ) framework.

[87] Dongarra et al., Applied Mathematics research for Exascale Computing, March 2014, Report, US Department of Energy, 2014

The following specific research topics are proposed:

1. Extreme-scale algorithms for forward in time computing: Implicit and semi-implicit algorithms for the numerical integration of non-linear dynamical systems have several decisive conceptual advantages over explicit schemes. These advantages include superior conservation properties, robustness for arbitrary time step sizes, and improved adherence to dominant physical balances. Implicit algorithms generally require more communication in parallel implementations, so that their advantages may be offset by diminishing efficiency on massively parallel future systems. Important open questions in this context which need to be addressed, concern the existence of algorithmic rearrangements that systematically improve the parallelism of implicit schemes, or alternatively the development of innovative approaches that would render explicit time discretisation competitive with implicit ones, both in terms of energy-to-solution and time-to-solution, as well as preserving desirable conservation and mimetic properties, robustness, improved fault tolerance to the propagation of local errors due to the tight processes coupling, and preserving fundamental balances, underlying the simulated dynamical systems. Specific attention should be given to solvers using higher-order integration schemes (such as Discontinuous Galerkin schemes) as these are needed, e.g., for industrial and engineering applications (see challenges WC1, chapter 4.3.1.2).

2. New approaches to highly scalable 3-dimensional Fast Fourier Transforms (FFT): The scalability of various exascale relevant applications depends on the availability of highly scalable FFT algorithms (see M2 and BM2 in Section 4.3.1). New approaches should aim for exploiting new network capabilities for performing collective communication operations and overlapping computations and communication.

3. Trading performance for accuracy through means of improved robustness: Robustness of algorithms becomes more important since it brings opportunities to trade performance for accuracy. Possible options are to use inexact, stochastic and hybrid, however, fast, communication-avoiding and communication-hiding methods for computing approximate solutions at intermediate steps or to reduce the impact of some global operations, such as reductions and/or dot-products, by decreasing their arithmetic accuracy or by skipping some irrelevant computations (see also **M-ENR-AR-7**).

Related milestones are **M-ALG-1 and M-ALG-2**.

6.7.2 Methods for scalable data analytics and artificial intelligence

Significant challenges arise from extreme data challenges, which require a paradigm shift from a compute centric view to a more data centric view. Algorithms for data discovery, and in particular those that discover global properties of data , such as graph analytics, require highly scalable compute resources and are not embarrassingly parallel. Accessing graph nodes across extremely large clusters may result in highly irregular memory and network access patterns and imposes a huge scaling challenge. Graph analytics are in particular necessary for discovering hidden connections and patterns within data and for identifying non-conforming objects in massive data sets (the needle-in-the-haystack problem). Other methods require analysis and visualisation to be embedded in highly-parallel simulations or real-time processing of massive data streams for event detection and support of ad-hoc decision making. Such operational and mission-critical applications in data analytics require specially balanced computer architectures which have particular bottlenecks removed and are not necessarily available on the standard IT market, also including new server design, I/O subsystems, and storage technologies. Some industry consortia have started discussing blueprints of special system architectures which will match their needs in data analytics. Analytic approaches are, therefore, heavily constrained by technological limitations. Progress, thus, has to be achieved through co-design activities involving both, designers of architectures and technologies, as well as developers of new mathematical methods and algorithms.

The following specific research and development topics are proposed:

1. Scalable analytics and artificial intelligence: Modern methods for data analytics and artificial intelligence often lack the ability to scale on massively-parallel HPC architectures. They can thus not efficiently exploit the capabilities of supercomputers whilst a growing number of use cases start to mandate supercomputing resources. This concerns, e.g. graph-based analytics featuring irregular

memory access and communication patterns. Further research is required to develop and establish methods, which will allow to scale-out problems on massively-parallel HPC architectures.

2. Enabling co-design of mathematical methods for data analytics and HPC technologies and architectures: In order to mitigate the technical limitations of today's analytic approaches on today's architectures, more efforts on co-designing mathematical methods for data analytics and application optimised HPC technologies and architectures are required. These efforts should result in a parametrisation of relevant data analytics use cases towards exascale performance levels.

Related milestones are **M-ALG-3** and **M-ALG-4**.

6.7.3 Mathematical support for data placement and data movement minimisation

As memory hierarchies of computers become increasingly complex and heterogeneous, the problem of how to decompose user data onto the various levels and how/when to move data between levels becomes an exponentially difficult problem. This problem is exacerbated by new memory technologies that support multiple usage models, and the introduction of data "persistence" (in non-volatile memories) as a new modelling requirement. The true complexity of the data problems being solved is not yet known, but seemingly simple data layout problems such as the tile-size selection problem, bear striking similarities to heterogeneous partition problems such as the MaxGrid problem, which can be shown to be NP-complete[88], as can many other static scheduling and partitioning problems. When inter-process communications and the effects of multiple data structures are combined, the complexity of data partition problems is likely to be NP-complete. If the general problems are shown to be of high complexity, then sensible heuristic, approximation or special-case problems must be defined instead.

The programming environment will seek to solve these data-related issues via additional language constructs, user controls as well as software- and/or hardware-based automation, e.g., of caching and pre-fetching. However, without advancements at the mathematical level, the achievable level of automation is expected to be low and thus the burden to the application development life cycle will increase. It is expected that a combination of new technologies will be required to support the data optimisation needs of the various classes of algorithms. However, currently, such algorithmic classes cannot be ascertained easily. Advances are required in such algorithmic categorisation, as well as in specific fields of I/O lower bounds, static (affine) scheduling algorithms[89] and dynamic (task-level) scheduling.

The following specific research and development topics are to be addressed:

1. The various classes of data, partitioning and scheduling problems and the classes of data movement problem that stem from key applications should be studied and categorised.

2. The complexity of the data problems should be ascertained, and specifically, the likely availability of a polynomial time solution to each category of problem, or alternatively, a list of suggested heuristic, approximation or special-case problems to solve.

3. Existing work on the defining of I/O lower bounds[90], which is the most appropriate measure of complexity for data problems, should be extended and deepened. The end goal of such work is to define a framework that can be adopted by programming environments so that data optimisations can be placed in context to the I/O lower bound. Ultimately, I/O lower bounds for arbitrary loop nests should be derived.

4. Mathematical support for static scheduling of codes should be developed in the areas of Lattice Optimisation, integer linear programming and other mathematical solvers, for solving data-driven scheduling problems at run-time.

5. Mathematical and algorithmic approaches for the scheduling of tasks on abstract resources should be developed and deepened. The applicability of the described approaches should be clear and the conditions under which the approaches are recommended should be equally clear. The approaches should extend to scheduling of problems over deep and heterogeneous memory hierarchies.

6. The mathematical support for the optimisation of multiple memory levels must be simultaneously developed.

[88] Casanova, Henri, Arnaud Legrand, and Yves Robert, 'Parallel algorithms', CRC Press, 2011.

[89] Bastoul, Cédric, et al. 'Putting polyhedral loop transformations to work' Languages and Compilers for Parallel Computing. Springer Berlin Heidelberg, 2004. 209-225.

[90] Irony, Dror, Sivan Toledo, and Alexander Tiskin. 'Communication lower bounds for distributed-memory matrix multiplication', Journal of Parallel and Distributed Computing 64.9 (2004): 1017-1026.

Related milestones are **M-ALG-5** and **M-ALG-6**. For reaching these milestones the architecture roadmaps for the next couple of years need to be assessed in respect to data storage and movement capabilities, including e.g. node-level memory hierarchy organisation. The outcome of this research is expected to impact work in programming environments.

6.7.4 How can the algorithmic and mathematical advances be leveraged in programming tools and resource schedulers

The assumption in this section is that new mathematical and algorithmic work is required before tools are developed, and that without such theoretical advances an incorrect or incomplete toolset could be developed. As such, it is imperative that the work described herein, takes a technology-neutral stance and can be leveraged by multiple toolsets.

There are, for instance, several areas of the programming environment that would be boosted by the mathematical and algorithmic advances stated above. New programming languages and their compilers could use the data movement and minimisation algorithms directly. This would involve incorporating the proposed models into a form that the compiler's internal representation could understand, and that the compiler can execute the set of code transformations that would be recommended by the solutions to the data problem. These same changes can be leveraged in run-times, but only in combination with some higher-level abstractions since the problem requires context which is not available in typical run-time libraries. The proposed changes also potentially support the development of tools that would not fit into the current standard environment. In particular, domain-specific languages could be designed specifically to support this data-centric analysis and code transformation. In all three of the mentioned cases, additional infrastructure work would be required that constructs the models, configures optimisation problems, and interfaces to the programming environment software.

Another area that requires further research in mathematical methods and algorithms is resource scheduling. Architectures are becoming more complex and heterogeneous, resulting in scheduling problems to become more challenging. Resource schedulers will additionally have to cope with dynamic changes of the requested resources and facilitate efficient system utilisation through new load balancing algorithms.

The following specific research and development topics are proposed:

- Mathematical methods for compiler technologies, run-time environments and related tools: Leverage results from research and developments related to data movement minimisation, to enable new compiler technologies that support optimisation of data placement and minimisation of data movement. This will in particular require a formulation of the proposed mathematical methods in terms of internal representations used by compilers. These mathematical methods will also impact run-time environments and tools to enable optimal data placement and data movement minimisation at run-time. This work should result in specific models and formulation of optimisation problems which can be used in compilers, run-time environments and related tools.

- New scheduling algorithms and techniques: Formulate new algorithms for scheduling of different types of resources provided by possibly heterogeneous HPC architectures including attached storage systems. These algorithms will have to be flexible enough to tackle the running application particularities, whose variety may imply the need a kind of metric to build off/on-line metrics for scheduling-level decisions. Explore the usability and effectiveness of the proposed approaches through integration in widely used resource schedulers.

Related milestone: **M-ALG-7**.

6.7.5 Algorithms reducing energy-to-solution

The potential of reducing energy-to-solution has been demonstrated for a number of cases where different methods for solving the same problem, revealed significant differences in terms of an energy-to-solution metric[91]. Minimisation of data movement is an important aspect in this context as (off-chip) data transport is the most expensive in terms of energy consumption. But, also other aspects need to be addressed, including effective exploitation of a given hardware architectures and the design of algorithms suitable for architectures, based on particular power-efficient technologies. Architectures are likely to become more heterogeneous to improve power efficiency at hardware level. A further aspect concerns improving our understanding of the interplay between

[91] Pavel Klavík, A. Cristiano I. Malossi, Costas Bekas, Alessandro Curioni, "Changing computing paradigms towards power efficiency", Philosophical Transactions A, 2014.

algorithmic choices and the energy required to solve a particular numerical problem. HPC solutions at various levels are becoming capable of providing fine-grained information on power consumption, and thus, facilitate energy-to-solution measurements. Generalised, designing algorithms that minimise energy-to-solution and changing computing paradigms towards a still to be established energy efficiency paradigm is the next step. This is further addressed in Chapter 5.4 **[M-ENR-MS-1]**. Furthermore, these energy measurement capabilities open opportunities for reducing energy-to-solution through auto-tuning frameworks. Research is required to better understand how different energy-efficient algorithms work together, in particular, in those cases where different data layouts are required.

The following specific research and development topics are proposed:

- Development of algorithms optimised for energy-to-solution: New algorithms should be developed which allow for a reduction of energy-to-solution on existing, emerging and future architectures. This research should include a characterisation and comparison of algorithms with respect to energy consumption (see also Chapter 5.4, **M-ENR-AR-8**).

Related milestone: **M-ALG-8**.

6.7.6 Vertical integration and validation of mathematical methods and algorithms

Efforts are required to ensure for any of the developed mathematical methods and algorithms that these can be efficiently implemented on different types of HPC architectures and can easily be used by a broad user community. This requires a cross-cutting approach together with other areas of the SRA to:

- Investigate the scalability of particular mathematical methods and algorithms at all relevant levels for relevant use cases, and work-loads and extrapolate scalability for upcoming exascale architectures;

- Explore and tune algorithmic parameters to maximise scalability and performance of the algorithms;

- Improve agility in the design, implementation or porting, tuning and optimisation process for different algorithms on different architectures.

Vertical integrations should ensure scalability at all levels, i.e. at mathematical models/methods level, through algorithmic level, down to systems and architecture levels. The interest is in an integrated approach in developing scalable mathematical methods and algorithms that lead to scalable programming models and tools and high-performance libraries optimised for specific architectures, all these ensuring scalability at all levels and opening a path to exascale scalability[92].

A suitable balance between hardware architectural and algorithmic performance needs to be identified as only the combination of both determines the overall performance in terms of minimal time-to-solution. This will also require a suitable tuning of algorithmic parameters support by auto-tuning techniques.

The results of such efforts can only be exploited if a careful design process for next generation mathematical algorithms is in place that is aware of the relevant roadmaps for HPC systems architectures. This becomes critical, in particular, for leveraging the potential of extreme scale algorithms when ported on the emerging heterogeneous computing architectures.

The following specific research and development topics are proposed:

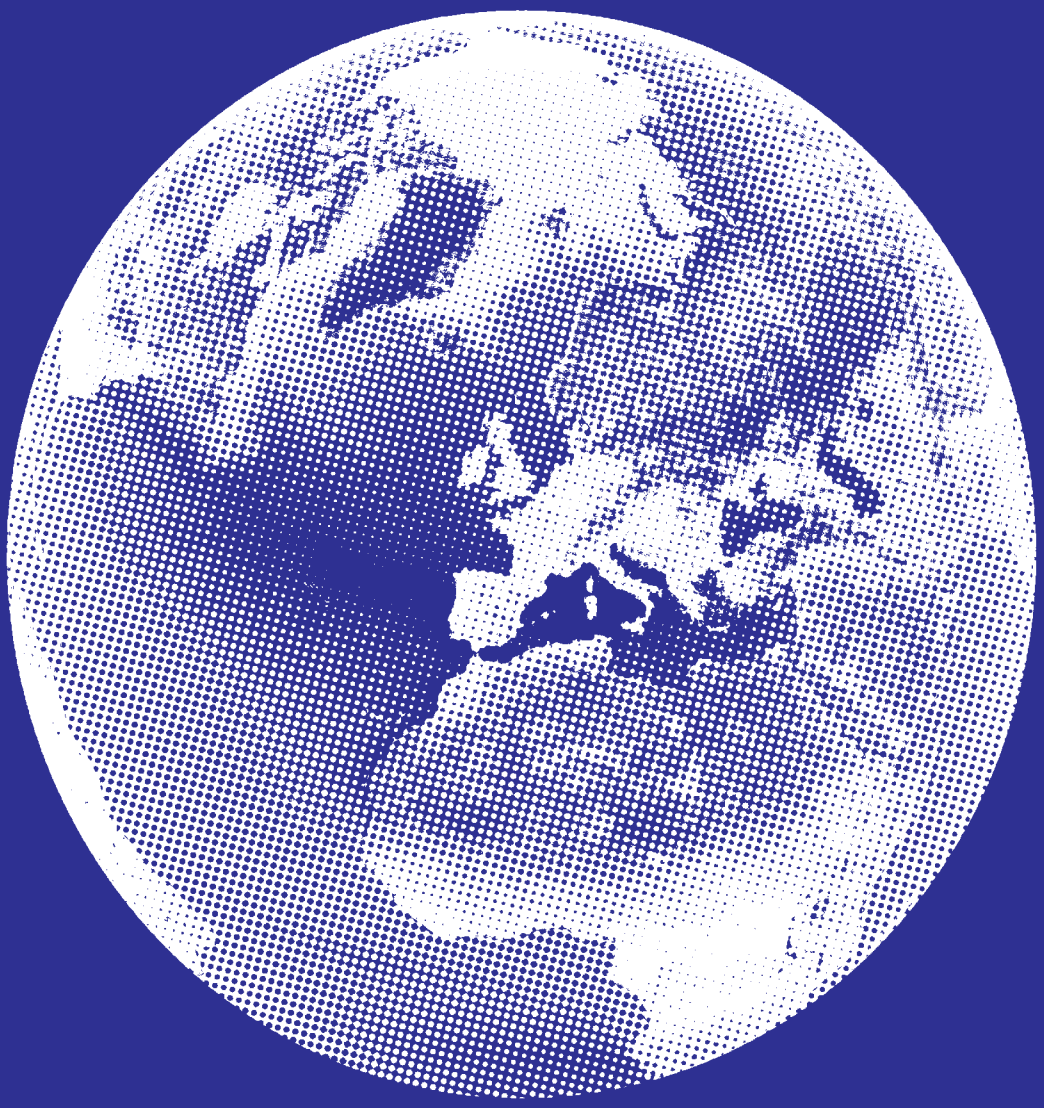
- Vertical integration and validation: Algorithms and mathematical methods are to be tested and validated with respect to scalability at all levels as well as ease of implementation, tuning and optimisation on different architectures. This must involve exploration of full vertical integration from lower levels of system hardware to upper levels of system software architectures.

- Tuning of algorithmic parameters for exascale: The parameter space for tuning algorithms to maximise their scalability, performance and energy efficiency on current, emerging, as well as future exascale architectures should be investigated.

Related milestones are: **M-ALG-9** and **M-ALG-10**.

[92] Alexandrov V. 'Scalable Stochastic and Hybrid Methods and Algorithms for Extreme Scale Computing', Procedia Computer Science, V. 29, Volume 29, pp. 1888–1892, Elsevier 2014.

7.



RESEARCH MILESTONES

A milestone in the tables below is defined as a tangible **research subject (content)** achieved by a **certain point in time** ('availability date'). The results of the research topic referenced by a milestone should be available at the availability date and might be needed to start work on some other milestone(s) in the same or another of the seven domains.

A milestone in a particular domain can have two relationships with one or more milestones in different domains:

- A milestone can be a **co-requisite** of one or several milestones of other domains (i.e. they all have the same or a "close-by" availability date). These milestones should be worked on in unison.

- A milestone can have one or more milestone from other domains as **pre-requisites**. This means that the results of these other milestones are needed in order to work on this specific milestone.

Wherever applicable, milestones are linked to requirements listed and explained in the "Application Requirements" (chapter 4.3). The column "Application Req." shows application requirements the results of a given milestone are supposed to fulfil.

7.1
HPC SYSTEM ARCHITECTURE AND COMPONENTS

SRA-3 SYSTEM ARCHITECTURE AND COMPONENTS MILESTONES	AVAILABILITY DATE	CO- REQUISITES	PRE- REQUISITES	APPLICATION REQUIREMENT
M-ARCH-1: New standard interfaces available for integrating CPUs and accelerators on nodes and to accommodate innovative unified memory and storage architectures on networks	2021			
M-ARCH-2 Having well balanced systems taking benefit from high bandwidth memories and NV memories	2019			
M-ARCH-3: Faster end-to-end communication networks (2x and 4x bandwidth in 2018 and 2021 compared to 2015 and lower latency) with energy and power used being proportional to bandwidth	2018, 2021			
M-ARCH-4: End to end optical communication chain including photonic switching in order to compensate network complexity growth on the larger fabric	2020			
M-ARCH-5: Optimised network and storage architectures available with dynamic features, QOS and virtualisation capabilities	2020	M-BIO-6		
M-ARCH-6 System and hardware to support performant direct remote memory access that would enable new and easier ways to program parallel applications	2020			
M-ARCH-7: Exascale system power envelope in the 5-15 MW power envelop range	2023-24			
M-ARCH-8: Exascale system available, at 100x more performance for relevant applications compared to today's state-of-the-art PRACE Tier-0 systems.	2023-24			

7.2
SYSTEM SOFTWARE AND MANAGEMENT

SRA-3 SYSTEM SOFTWARE AND MANAGEMENT MILESTONES	AVAILABILITY DATE	CO-REQUISITES	PRE-REQUISITES	APPLICATION REQUIREMENT
M-SYS-OS-1 Memory Hierarchy-management policies and libraries for NVRAM	2019	M-BDUM-PROG-2 M-BDUM-MEM-2	M-ARCH-2, M-ARCH-3, M-ALG-5	A1
M-SYS-OS-2 OS decomposition and specialised containerisation	2019	M-BDUM-VIRT1 M-BDUM-VIRT-2	M-ARCH-8	FS2
M-SYS-OS-3 HW Embedded Security integration and cross layer security support	2018-2021			
M-SYS-IC-1 Efficient peer to peer and storage over fabrics support	2019/2020	M-PROG-1 M-BDUM-MEM-2	M-ARCH-1 M-ARCH-4 M-ARCH-5	
M-SYS-CL-1 Initial support of mixing HPDA, AI and HPC environment	2018/2019			
M-SYS-RM-1 Resource management and orchestration support for complex workflow	2018	M-PROG-4		M1, GSS2
M-SYS-RM-2: data aware and Multi-criteria resource allocation integration for adaptive scheduling	2019	M-BDUM-PROG-2 M-BDUM-VIRT-2		
M-SYS-RM-3 Dynamic reconfiguration scheduling support (for flexibility and resiliency) purpose	2019	M-ALG-7 M-PROG-4		
M-SYS-RM-4 Data aware and power efficient scheduling	2020	M-ALG-6 M-ALG-7 M-PROG-1 M-ENR-MS-3	M-ALG-8	BM1
M-SYS-VIS-1 software support for In-situ computation and visualisation	2019			FS1

7.3
PROGRAMMING ENVIRONMENT

SRA-3 SYSTEM ARCHITECTURE AND COMPONENTS MILESTONES	AVAILABILITY DATE	CO- REQUISITES	PRE- REQUISITES	APPLICATION REQUIREMENT
M-PROG- 1: APIs and corresponding libraries, run-time and compiler support for auto-tuning of application performance (incl. energy use) and supporting legacy codes.	2020	M-ALG-[5,6,7,10]; M-BIO-3; M-ENR-[MS1, MS2]; M-SYS-[IC-1, RM4]	M-SYS-OS-1	A1 A3 M3
M-PROG- 2: High-level programming and domain-specific language frameworks	2020			A3
M-PROG- 3: Non-conventional parallel programming approaches (i.e. not MPI, not OpenMP / pthread / PGAS - but targeting asynchronous models, data flow, functional programming, model based).	2021	M-BDUM-PROG-3, M-ALG-5	M-BIO-5	
M-PROG-4: Enhanced programming model and run-time system support for dynamic environments (management & monitoring), optimisation of communication and data management, interaction with OS or VM - within application workflows.	2020	M-BDUM-VIRT-1; M-ENR-[MS-3, FT-7]; M-SYS-[RM-1, RM-3];		
M-PROG-6: Performance Analytics and Debugging tools at extreme scale, including data race condition detection tools and user-support for problem resolution.	2021			A3
M-PROG-7: Performance analytics and debugging tools co-designed to link to the application developer's original code and high-level programming environments	2022			

7.4
ENERGY AND RESILIENCY

SRA3 ENERGY AND RESILIENCY MILESTONES	AVAILABILITY DATE	CO-REQUISITES	PRE-REQUISITES	APPLICATION REQUIREMENT
M-ENR-MS-1: Characterisation of computational advance as function of energy/power metric and standardisation of this approach with automatic and semi-automatic tools	2019	M-PROG-1		
M-**ENR-MS-2**: Methods to manage computational advance based on the pre-set energy/power metric and achieve Proportionality Computing with respect to the selected metric	2020	M-PROG-1 M-BIO-7	M-ENR-MS-1 M-BIO-7	WC2
M-ENR-MS-3: Throughput efficiency increase by scheduling instructions to the cores and functional units in the processor within its power envelope and taking the time criticality of the instructions into account	2022	M-SYS-RM-4 M-PROG-4 M-BIO-5	M-SYS-RM-3	
M-ENR-HR-4: Optimisation of the energy spend by the facility by controlling the coolant temperature down to the device level and taking the infrastructure energy cost into account	2021	M-BIO-3	M-ARCH-1 M-ARCH-2 M-ENR-FT-5	
M-ENR-FT-5: Collection and Analysis of data from sensor networks - the Big Data challenge for measurements around the facility	2020	M-BIO-4		
M-ENR-FT-6: Prediction of failures and fault prediction algorithms	2021	M-BIO-2	M-ENR-FT-5	
M-ENR-FT-7: Application recovery from fault conditions in the system	2022	M-ENR-FT-5 M-ENR-FT-6 M-PROG-4		
M-ENR-AR-8: Energy/Power efficient numerical libraries	2020		M-ALG-8	
M-ENR-MS-9: Highly efficient HPC installation	2021	M-ENR-MS-1 M-ENR-MS-2 M-ENR-MS-3	M-ENR-HR-4 M-ENR-FT-5	

7.5
BALANCE COMPUTE, I/O AND STORAGE PERFORMANCE

SRA-3 I/O AND STORAGE MILESTONES	AVAILABILITY DATE	CO- REQUISITES	PRE- REQUISITES	APPLICATION REQUIREMENT
M-BIO-1: One or more storage class memory technology usages demonstrated as part of the persistent storage hierarchy.	2019		M-SYS-OS-1	GSS1, GSS2 A4,
M-BIO-2: Extreme scale storage and I/O system simulation framework established.	2020	M-ENR-FT-6		
M-BIO-3: Standardised Extreme scale I/O middleware API available: incorporating advanced features such as data layouts on NVRAM/Flash/Disk, in-storage computing, Object stores, etc,and also portability concerns raised by the CoEs.	2022	M-PROG-1 M-BDUM-MEM-2 M-ENR-HR-4	M-SYS-OS-1	A1
M-BIO-4: Big Data analytics tools developed and optimised for Storage and I/O.	2019	M-BDUM-MEM-2 M-BDUM-DIFFUSIVE-1 M-ENR-FT-5	M-SYS-CL-1	
M-BIO-5: In-storage compute capability across all tiers/layers of the storage system as indicated by the data requirements within the CoEs.	2020	M-BDUM-DIFFUSIVE-1 M-ENR-MS-3		WC2
M-BIO-6: I/O Quality-of-Service capability available for extreme scale storage systems.	2020	M-ARCH-5	M-SYS-RM-4	
M-BIO-7: Extreme scale multi-tier data management tools available.	2019	M-ENR-MS-2	M-SYS-OS-1	
M-BIO-8: Completion of co-Design with new use cases identified by the CoEs (AI/Deep learning, etc.)	2019	M-BDUM-ALGS-1	M-ALG-3	

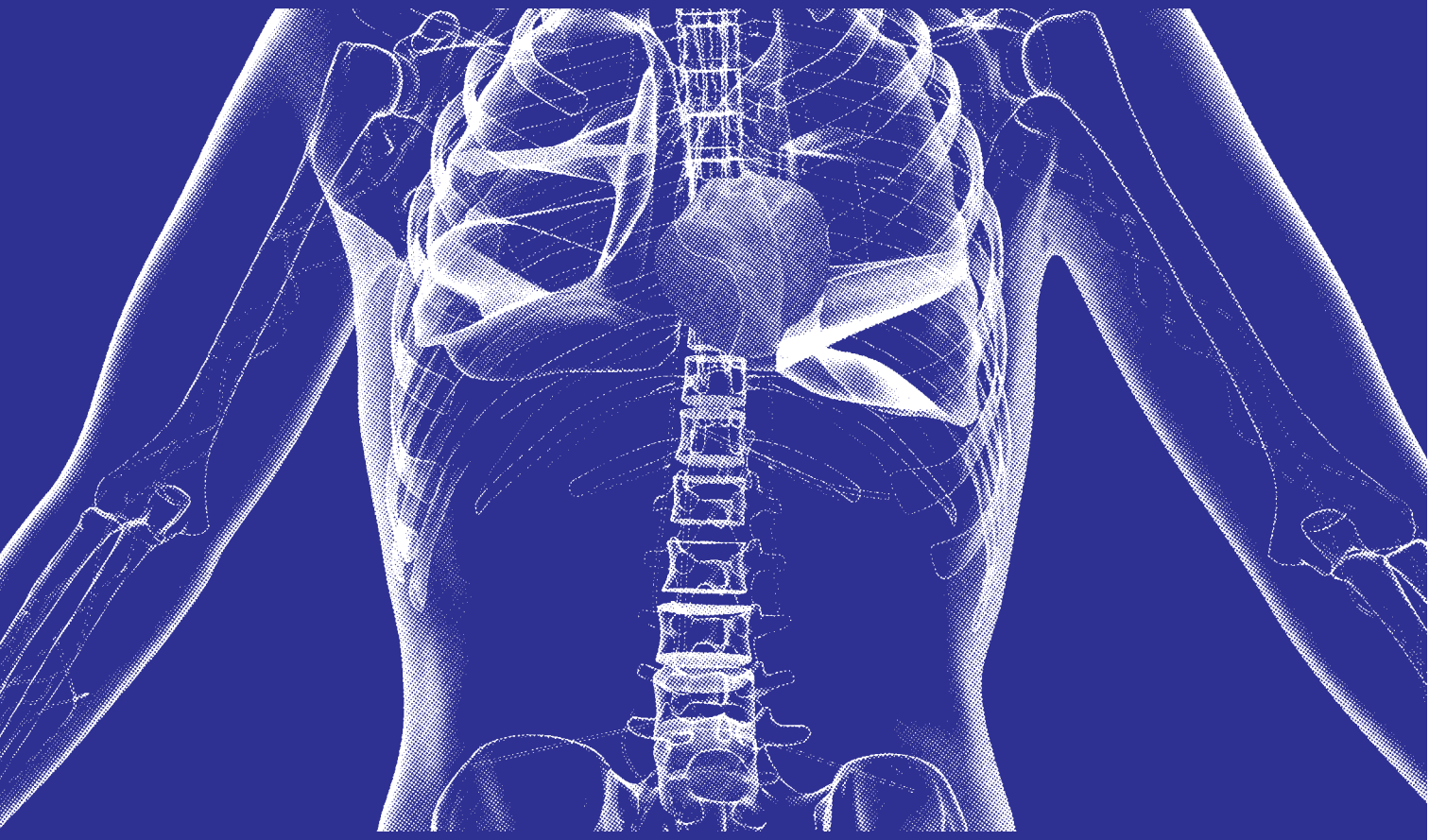
7.6
BIG DATA AND HPC USAGE MODELS

SRA3-BIG DATA AND HPC USAGE MODES MILESTONES	AVAILABILITY DATE	CO-REQUISITES	PRE-REQUISITES	APPLICATION REQUIREMENT
M-BDUM-METRICS-1: Data movement aware performance metrics available.	2020			
M-BDUM-METRICS-2: HPC-like performance metrics for Big Data systems available.	2020			
M-BDUM-METRICS-3: HPC-Big Data combined performance metrics available	2021			
M-BDUM-MEM-1: Holistic HPC-Big Data memory models available.	2020		M-SYS-OS2	
M-BDUM-MEM-2: NVM-HPC memory and Big Data coherence protocols and APIs available.	2021	M-SYS-OS1, M-SYS-IC1 M-BIO-3 M-BIO-4		
M-BDUM-ALGS-1: Berkeley Dwarfs determination for Big Data applications available.	2020	M-BIO-8		
M-BDUM-ALGS-2: Dwarfs in Big Data platforms implemented.	2022	M-ALG-4		
M-BDUM-PROG-1: Heterogeneous programming paradigms for HPC-Big Data available.	2020			
M-BDUM-PROG-2: Heterogeneous programming paradigm with coherent memory and compute unified with Big Data programming environments available.	2021	M-SYS-OS-1 M-SYS-CL-1		
M-BDUM-PROG-3: Single programming paradigm across a hybrid HPC-Big Data system available.	2023	M-PROG-3		
M-BDUM-VIRT-1: Elastic HPC deployment implemented.	2021	M-PROG-4 M-SYS-OS-2		
M-BDUM-VIRT-2: Full virtualisation of HPC usage implemented.	2023	M-SYS-CL-1 M-SYS-OS-2	M-ARCH-8	
M-BDUM-DIFFUSIVE-1: Big Data - HPC hybrid prototype available.	2020	M-BIO-4 M-BIO-5	M-BIO-5	
M-BDUM-DIFFUSIVE-2: Big Data - HPC large-scale demonstrator integrated.	2023		M-SYS-CL1	

7.7
MATHEMATICS AND ALGORITHMS FOR EXTREME SCALE HPC SYSTEMS

SRA3 MATHEMATICS AND ALGORITHMS MILESTONES	AVAILABILITY DATE	CO- REQUISITES	PRE- REQUISITES	APPLICATION REQUIREMENT
M-ALG-1: Scalability of algorithms demonstrated for forward in time computing and 3-dimensional FFT for current architectures.	2019			E2, BM2, M2, 1IND1
M-ALG-2: Multiple relevant use cases demonstrated for improving performance by means of robust, inexact algorithms with reduced communication costs.	2018			
M-ALG-3: Scalable algorithms demonstrated for relevant data analytics and artificial intelligence methods.	2019		M-SYS-CL-1	
M-ALG-4: Processes established for co-design of mathematical methods for data analytics and of HPC technologies/architectures.	2019	M-BDUM-ALGS-2	M - B D U M - ALGS-1	
M-ALG-5: Classes of data, partitioning and scheduling problems categorised and their complexity ascertained.	2019	M-PROG-1 M-PROG-3		
M-ALG-6: Mathematical and algorithmic approaches established for the scheduling of tasks on abstract resources and exploitation of multiple memory levels.	2019	M-PROG-1 M-SYS-RM-4		
M-ALG-7: Research on mathematical methods and algorithms exploited for compiler technologies, run-time environments, resource schedulers and related tools.	2020	M-PROG-1 M-SYS-RM-3 M-SYS-RM-4		
M-ALG-8: Reduction of energy-to-solution demonstrated by means of appropriately optimised algorithms demonstrated for a set of relevant use cases.	2018			
M-ALG-9: Process for vertical integration of algorithms established together with the validation of scalability, ease of implementation, tuning and optimisation.	2018			
M-ALG-10: Tuning of algorithmic parameters at exascale completed for a relevant set of algorithms.	2020	M-PROG-1		

8.



EXTREME-SCALE DEMONSTRATORS

High-Performance Computing (HPC) is a crucial asset for driving Europe's product innovations and stretching its technology providers. The "Extreme-Scale Demonstrators" (EsDs) are vehicles to optimise and synergise the effectiveness of the entire HPC H2020 Programme through the integration of isolated R&D outcomes into fully integrated HPC system prototypes; a key step towards establishing European exascale capabilities and solutions. The primary focus of the EsD projects will be establishing proof-points for the readiness, usability and scalability potential of the successful technologies developed in previous European actions (FP7, WP2014/15 and WP2016/17 and other R&D in Europe), when deployed in conjunction with open market technologies at that time.

There is an existing consensus between HPC centres and industrial members of the ETP4HPC that such projects should create "ready to use" systems commensurate with exascale commercial objectives. They should encourage a strong co-design approach between technology and applications providers. They would produce tangible results to validate the capabilities produced in the preceding H2020 work programmes. These EsDs should provide platforms deployed by HPC centres and used by the involved application providers for their production of new and relevant applications. Hence, the EsD should have an up-time compatible with the development and execution of applications, which also implies that applications by way of co-design need to be advanced to a readiness level that allows exploiting the technological advancements EsD are expected to provide. Applications in this context comprise existing and emerging use cases that require substantial performance boosts to address key science challenges with high relevance to European society and economy.

The fully integrated EsD systems should not be confused with systems/subsystems prototyped as part of individual research projects but as synergetic and integrated hardware and software platforms. At project end, the EsDs will have a higher TRL (Technical Readiness Level) of 7-8 (compared to 6-7 of prototypes as part of the projects), thus, their stability and usability will enable stable application production at reasonable scale . Therefore, the EsDs will be 'stepping stones' towards a more expedited and solid commercial exploitation of the underlying system design and technology. A clear path towards subsequent commercial exploitation of EsD output in preparing exascale level products and/ or component technologies must be provided. Whilst for the integration of EsDs

the target is to deploy technology developed in the FETHPC programmes, the EsD projects need to be open to also include relevant technology developed outside of this programme, i.e. in the other parts of the H2020, FP7 and member state programmes or in the global market.

To be clear, the purpose of creating and using the EsDs is fundamentally different from procuring 'big commercially available production systems' e.g. as done Tier-0 centres within PRACE. The EsDs are meant to validate and prove the advancements in R&D performed within the FP7 and H2020 HPC work programmes and gather valuable feedback for future projects, whilst the periodically procured commercial HPC systems are geared towards providing a robust compute infrastructure available to large user communities. However, a fraction of cycles from the EsDs, once deployed and stable, should be made available to members of larger user groups, with well-established allocation mechanisms to also expose the technology to the wider community. The EsD projects should, therefore, help industrial (and also SME and Mid-Cap) users to prepare for the next step in their HPC usage, by being efficient not only on the existing applications, but also on the emerging ones.

In summary, the EsD projects will fill the following important gaps in the current HPC H2020 programme:

- Bringing together the technologies developed in the FETHPC programme and related H2020 R&I and boosting them closer to commercialisation, thus, fostering exploitation and take-up of these technologies

- Benefiting from targeted R&D efforts across many projects and combine components into an integrated system

- Providing the missing link between the three pillars of the HPC strategy: technology providers, infrastructure providers, and user communities (application providers and users) through projects that leverage their respective expertise to develop new high-end compute platforms

- Enable existing and emerging applications to demonstrate much enhanced performance given well defined, application specific targets as well as productive use of the new architecture and technologies

8.1 PHASES OF ESD PROJECTS

EsD projects will address the design, integration, validation and experimentation of systems capable of being commercialised, operating in a close-to production mode and leveraging results of the technologies developed in projects funded by FP7 or H2020 or other R&D actions in Europe. Each project will follow a 2-phase approach:

Phase (A): Development, Integration and Testing, involving applied technology research projects, which will have a substantial R&D focus mostly geared towards integrating and customising hardware and software components and sub-systems developed in the preceding R&D projects. Performance requirements will be defined in function of applications

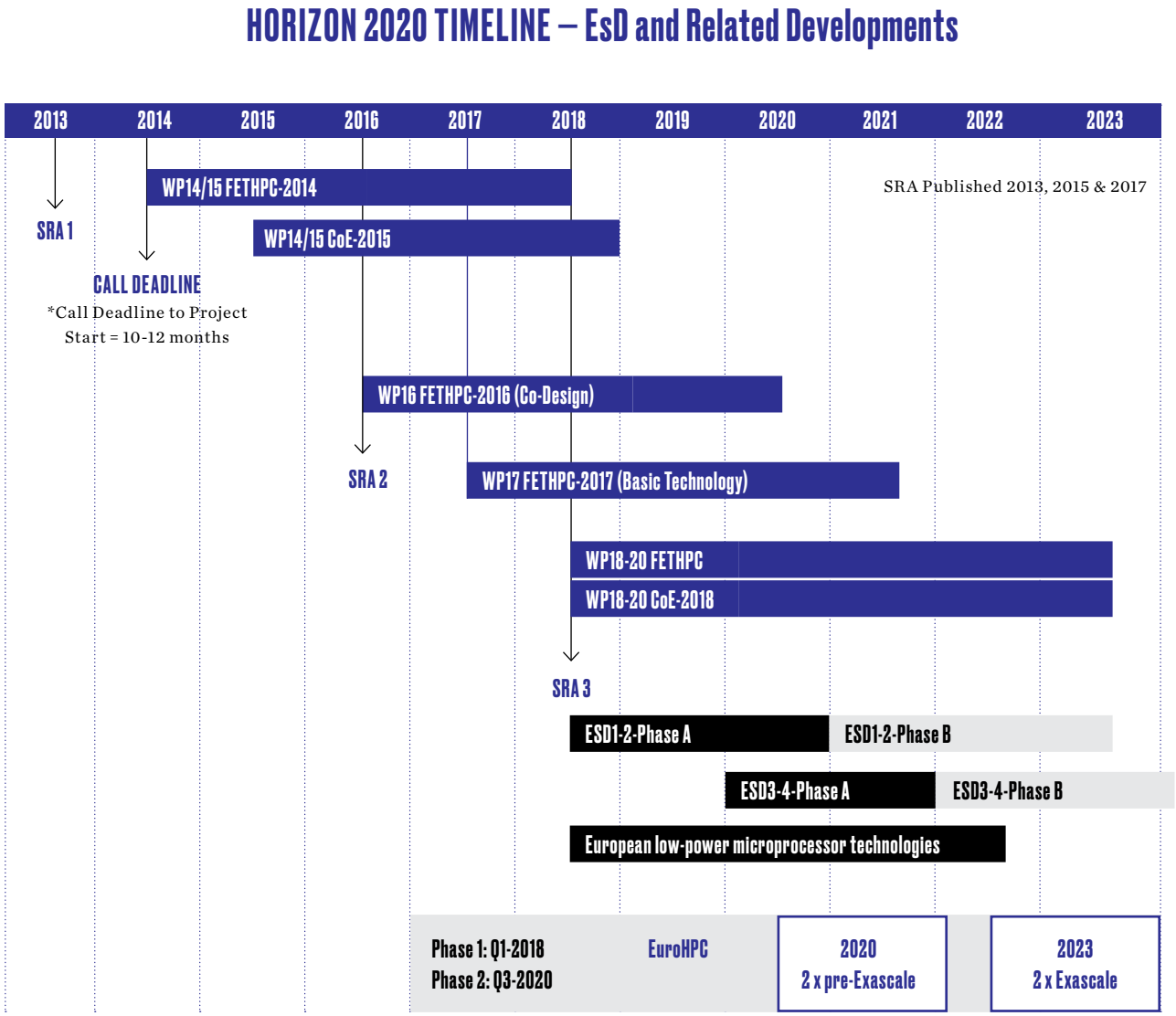
Phase (B): Deployment and Use, where the EsD is validated and operated by a hosting centre and made available to application owners for code porting and development to address numerical/extreme data challenges as well as characterisation and platform validation based on real use cases

Phase A will lead to the development of a system that is fully usable by the end of this phase, with the deployment of a system of sufficient size to enable evaluation and validation of the design. The EsDs should be the result of co-design between applications on one hand and hardware plus system software on the other. Achieving well specified performance targets with a representative set of applications in both phase A (in demonstration mode) and B (in production mode) is considered key for a successful EsD. Sufficient resources must be made available to ensure porting and adaptation of applications to each EsD, also allowing room for research and development towards new programming models, and possibly, algorithmic approaches being enabled by new technologies.

8.2
SCOPE OF ESD PROJECTS

Given the main purpose of the EsDs to bring results from European R&D&I on HPC closer to commercialisation, it is key that EsD projects become a continuum rather than a one-time effort. Nevertheless, to maximise impact of the EsDs, the timing and concrete objectives of EsDs should be aligned with the European HPC agenda. This includes the FETHPC Work Programmes, but also other important milestones such as the envisaged European Exascale Systems and European Processor initiative as well as related member state initiatives such as EuroHPC and IPCEI.

Figure 14.
The European EsD developments mapped onto the timeline of the European Horizon 2020 research time-line.



In a first set of EsD projects starting in 2018/2019 results from WP14/15, FP7 or other R&D in Europe should come to fruition with a primary target to demonstrate an achievable energy efficient and scalable system-level technology path to commercially viable and competitive Pre-Exascale systems by 2021/2022 and Exascale systems by 2023/24. The demonstration of scalability up to Exascale levels (i.e. a design point enabling a performance level that enables about 100x more performance compared to today's PRACE Tier-0 systems) is an important criterion. It needs to be stressed that progress in full precision floating point performance does not necessarily have to be the only dimension to demonstrate progress towards exascale. Other areas and related milestones identified in the SRA, such as architecture, components, I/O and storage, big data, are relevant as well. The resulting EsDs should be as equilibrated as possible in the various dimensions in order to get good performances/W/€, but should clearly contribute to pre-exascale systems with energy-efficiency and lowering the cost of ownership in mind.

Phase B of projects in this call will allow to evaluate and validate the complete system. It must support real applications aiming at relevant scientific and/or economic communities and applications, with a potential to exploit early Exascale class systems. The system must demonstrate operational and performance characteristics suited to a number of key characteristic usage models (one size may not fit all) in its areas of application for high-end world-class HPC use by 2021/22 and should target commercial viability.

EsDs in 2020 and beyond should additionally build upon results from WP16/17 and later, and reflect the latest technological advancements, including the European processor initiative. If possible, it should also learn from the first EsD phase. Furthermore, new deployment areas for HPC system infrastructure should be explored and demonstrated such as (however, not limited to) High Performance Data Analytics (HPDA) and Machine Learning (ML) or others. Exploiting advantages offered by novel mathematics, such as transmathematics or Extreme compute, is still a valid subject field if sufficiently differentiated from the state of the art.

Consequently, Phase B for such future EsDs will enable new application areas such as (however, not limited to) High Performance Big Data Analytics (HPDA), Artificial Intelligence (AI), a combination of them or others areas. The user communities must demonstrate operational and performance excellence in a number of key characteristic usage models for high-end world-class HPC use with expected deployment in 2024 time frame. The system should also demonstrate economic viability (cost of ownership, energy efficiency). In general, EsDs should also be used to address challenging scientific problems in Phase B, for which EsD technology promises breakthrough performance. This will not only be the final proof point for their maturity, but also leverage the substantial investment in building and operating these systems and into porting the applications.

The EsD calls will have a high dependency on the outcome of WP2014/15 and WP2016/17 projects regarding their timing, but mostly regarding contents. The portfolio of accepted projects in the work programmes must provide a sound technology basis for building EsDs, and the accepted projects should be actively encouraged to foster cross-project interlock. The structure of the call should support cross-project integration, with particular regard to IP visibility and licensing clarity. Little coherence between accepted projects, too many disjoint focus areas, and insufficient technology options and readiness might otherwise jeopardise the success of the EsD calls.

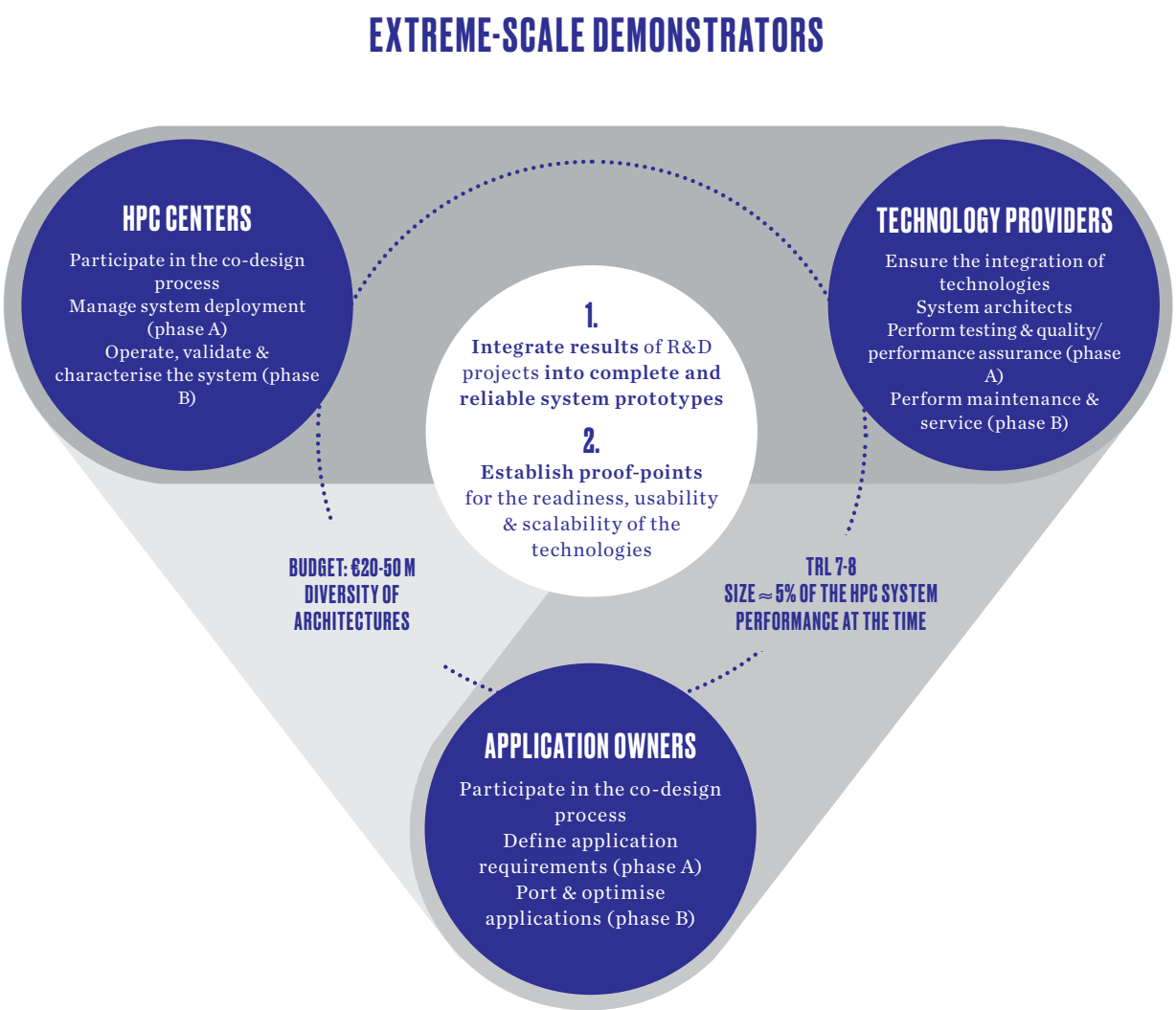
ETP4HPC proposes that the EsD project calls will have a funding envelope compatible with a spending of €20- 40M[93] (30-50% of the potential funding committed should be dedicated to R&D and 50-70% to the cost of the parts costsused) per EsD project for phase A and €3-6M for phase B to cover utilities, operation-manpower and maintenance. Each call should fund between one and two EsD projects, according to the available budget and the co-funding structure. Phase A should have a duration of 18-24 months and phase B of 24 months with a validation feedback checkpoint after 9 months. Therefore, total project duration of 32-48 months is envisaged.

[93] As of November 2017, the EsDs are included in the EC Information and Communication Technologies part of Work Programme 2018-20 under ICT-14-2019 with the expected funding 20-40 million Euro per project, available at
http://ec.europa.eu/research/participants/data/ref/h2020/wp/2018-2020/main/h2020-wp1820-leit-ict_en.pdf

8.3
ETP4HPC'S PROPOSAL FOR ESD PROJECT STRUCTURE

ETP4HPC recommends suitable projects to involve three types of partners for EsD projects: technology providers, application owners and HPC centres to ensure that all essential competences are includes in the consortium.

Figure 15.
The role of the European HPC ecosystem stakeholders in EsD projects.



The role of technology providers (with a key role for system integrators) will be to ensure the integration of various technology (hardware and system software), the project management, the testing and quality/performance assurance during phase A. The system integration will be the focal point for maintenance and service during phase B.

It is acknowledged that the preparation of an EsD proposal will require several months (in terms of time & effort) due to the analysis of the integration of technologies from different projects and reaching a design point to a level that allows a feasibility status and a cost estimate. There is a large investment in building the systems and therefore it would be good (if this is possible at all) to reuse parts of demonstrators done during the previous projects.

The consortium should have defined a clear management of Intellectual Properties in view of potential commercialisation of the results of the project: EsDs must not be one-of systems, but open a product line, and the substantial investment of project partners needs to pay off; all these aspects should be preferably agreed before the start of the project.

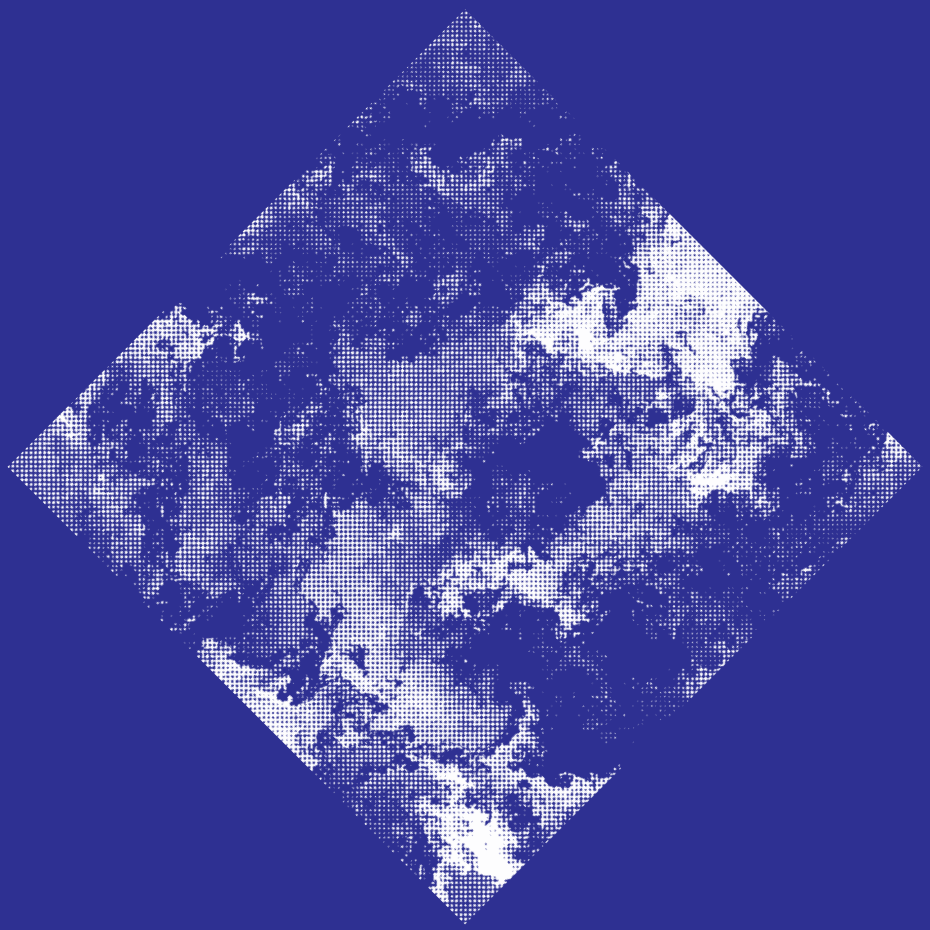
The role of the application owner will be to define application requirements and key challenges, which requires Tier-0 type resources, and that can be addressed by EsD during phase A. To enable a successful co-design, applications owners have to be involved from the very beginning of the project. It should be a true co-design cycle, with clear interactions between the user's requirements and the systems architects; none of the two taking the lead over the other. During phase B, they will port and optimise application(s) to EsD and use EsD productively. The effort required in this phase will vary depending on the architecture of the EsD and can be substantial, if it includes disruptive hardware- and/or software technologies such as new programming models. Each EsD project should show a significant engagement for application porting and optimisation and ensure application owners being rewarded by means of significant compute resources that significantly exceeds those typically awarded through large-scale PRACE projects.

The role of participating HPC centres will be to participate in the co-design process and to manage system deployment during phase A. They will have also to contribute to the development and adaptation of system software and tools to the new platform in order to take into account and optimise usage of innovative components. The collaboration on tools and system software stack could imply key technology provider and open source communities. Furthermore, they will operate the EsD, ensure interoperability with PRACE Services, validate and characterise the system prototypes (in terms of performances, robustness, efficiency, etc.) during phase B. The consortium proposing an EsD project should propose a funding and acquisition model allowing a sustainability of the use of the EsD systems after the end of the funded project. The EsDs need to be stable enough to allow production runs, and therefore should show a clear view on programming models and software stack developments. ETP4HPC recommends supporting commonly used tools (compilers, MPI, OpenMP, numerical libraries, etc.) as much as possible, and novel technologies should provide standard interfaces where possible.

In summary, with the targeted high TRL and substantial capability, the EsDs will evolve into valuable HPC resources towards the end of phase B, and it is highly desirable that they will be operated beyond the end of the project. It is therefore mandatory that the consortia agree on a number of questions such as the ownership of the system after the end of the project, coverage of the operational costs, to whom and how access will is granted. These decisions will also have an impact on the selection of a suitable acquisition model for the EsD which will have to consider H2020 funding rules as well as procurement regulations. After completion of the first set of EsD projects, lessons learnt should be discussed between the ETP4HPC and the EC, possibly leading to recommendations for adaptations in future working programmes or the next Framework Programme.


PRE-REQ
CO-REQ

9.



NON-TECHNICAL RECOMMENDATIONS AND PRIORITIES

In addition to the technical research priorities detailed in the previous chapters, this chapter contains several recommendations and priorities relevant to improve and further develop the HPC ecosystem in Europe, the situation of SMEs and start-up companies on the technology provision side and the important area on education and training[94].

9.1 ECOSYSTEM-LEVEL HOLISTIC RECOMMENDATIONS

EXDCI has identified a set of recommendations for the entire European HPC Ecosystem. The following paragraph gives a short overview and addresses three domains:

- Better research instruments - The following recommendations aim at improving current research instruments, both computing resources and deployment of new technologies in order to better support applications and researcher discovery processes.

  > R1: Design a new operation policies and federations towards convergence.
  > R2: Reinforce Big Data and extreme scale international initiatives.
  > R3: Improved access to advanced technologies.

[94] also a work-package in the EXDCI project

· R&D efficiency - The following recommendations purpose is to ensure that the public and private investment in R&D is carried out in a coherent manner, maximising the impact of research.

> R4: IPCEI for advanced research and innovation.
> R5: Paving the way from EsD[95] development towards applications.
> R6: Improving FETHPC and CoE result capitalisation.

· Industry competitiveness - The following recommendations' ambition is to leverage R&D excellence and translating its output into industry competitiveness:

> R7: Encouraging commercial relationships between SMEs and industry via European projects
> R8: Concerted approach to HPC training in Europe
> R9: Incentives to increase EU stakeholders' implications in standard initiatives

They are complementary to the technical ones proposed in this SRASRA. Details can be found in EXDCI Deliverable 4.3 'First holistic vision and recommendations report[96]'.

In conclusion, we are facing a 'paradigm shift' driven by the data and computing convergence. It creates many opportunities but also requires the European HPC community to adapt and invent new methods and approaches in science and industry. This large-scale effort is the first step to bridge applications and technology while encouraging the dialog between all the stakeholders.

9.2 SMES AND START-UPS

SMEs continue to play an important role in the HPC Ecosystem.

SMEs and start-ups are key players as providers of hardware solutions, software solutions and integration activities generating an important source of innovative power in the industry. Also, they are at the centre of the commercialisation of research-results. Financial, regulatory and project-related hurdles prevent SMEs and start-ups from competing with the established players. HPC is an investment intensive playground, with a few large projects or initiatives leading the entire development. In order to succeed, small companies require the resolution of a number of issues.

9.2.1 Market restrictions and barriers

European SME's are still confronted with market barriers, which prevent medium sized enterprises to grow. Tenders in the public sector are often still restricted for service providers with a minimum size. Size is defined by number of FTEs, historical sales or balance sheet ratios. A track record is also often set as a precondition for participation in tenders. SMEs and start-ups, which have no proven track record of historical sales to, for example, Top500 organisations are not allowed to tender, even though they might be offering innovative and competitive solutions.

Another limiting factor is the lack of funding or unfavourable payment conditions of organisations in the public sector. These terms often include a final payment upon the full completion of the solution in question, without any consideration for the large amounts invested in delivering the product – a heavy burden on any small company's cash flow. Large hardware providers demonstrate a higher level of financial flexibility, while small companies cannot afford such terms.

Oftentimes, the reasons above cause European SMEs and start-ups to become vulnerable targets of buy-outs by foreign investors, reducing the pool of the indigenous European companies.

[95] Extreme-Scale Demonstrators, http://www.etp4hpc.eu/en/esds.html
[96] Available at: https://exdci.eu/sites/all/themes/exdci_theme/images/D4.3.pdf

9.2.2 Proposed actions

The ETP4HPC Working Group (WG) for SMEs has been very active over the last year and will continue to do so in the next years to continue to build a globally competitive European SME-HPC technology value chain in the coming years.

The SME workgroup is in the process to develop a long-term strategy, which will form an important role in the discussion with the EC (cPPP), research organisations and large HPC-providers.

In order to do so, the Working Group will organise a number of workshops for members of ETP4HPC, not only SMEs. Focus of these workshops will be the formulation of a common mission statement and the design of cross industry congruent goals to support and improve the position of European SMEs in the European HPC industry. The topics which have been identified during the last years, as described here below, are currently being addressed.

Market barriers as mentioned above, for example, having participated in a prior Top500 project or demonstrate a certain minimum size of the enterprise only prevent smaller SMEs or start-ups from participating in tenders, thus limiting further development of innovative ideas or solutions.

The "European SME instrument[97]" is a funding tool to support the development and the 'bring to market' of technical, innovative solutions, developed by European companies. Until 2016, only a few HPC oriented companies have applied for this source of funding, which also restricts fast and successful development of new solutions. The work group aims to promote this tool in the HPC-SME environment and co-design specific conditions with the European Union to improve accessibility for SMEs.

In order to further support early introduction of innovative ideas, SMEs, Research institutions and other users of HPC in the public sector should be motivated to cooperate on a structural level in order to co-develop new systems and test innovative solutions, before possibly integrating them into a final product. In this field, the ETP4HPC SME Working Group will work in close cooperation with the EXDCI initiative, combining strength and optimizing synergies between Start-ups and SMEs.

As mentioned above, SMEs, particularly in the HPC-industry are often confronted with substantial funding requirements, as public sector related organisations often do not create favourable payment conditions for them. Together with the European Union, ETP4HPC plans to design and promote new standards or guidelines for the industry, which should improve the position of SMEs.

Last, but not least, the European HPC ecosystem needs to provide a common platform for R&D organisations, European users of HPC, European providers of HPC solutions, be it SMEs or large European industrials and the European Union. This long-term strategy also includes the participation of Start-ups and SMEs in the development of Extreme Scale Demonstrators and the long-term realisation of the European policy to play a major role in the global HPC ecosystem.

The ETP4HPC SME Working Group supports the extension of the recommendations above onto all SME active in the HPC area, also those outside of ETP4HPC.

[97] http://ec.europa.eu/programmes/horizon2020/en/h2020-section/sme-instrument

9.3
EDUCATION AND TRAINING

In order to address the issues identified in the first Strategic Research Agenda, the European Extreme Data and Computing Initiative (EXDCI)[98] has focused on three key objectives:

· Supporting Talent Generation

> To start from the ground up, by raising awareness of what HPC is, and the variety of job opportunities to which HPC skills can lead;
> To break down preconceptions of the sort of people who work in this field and show that the workforce is young and diverse, and to encourage young people to consider HPC as a career.

· Identifying and Meeting Future Training Needs

> To help those who have decided they want (or might want) to work in HPC to find the most suitable training opportunities to help them prepare for a career in HPC, whether this is through self-paced learning or attending courses or summer schools, whether face-to-face or on-line, and regardless of location;
> To help identify the training needs of the future and how best to meet these ever-changing needs.

· Facilitating HPC Staff Recruitment

> To help those who have gained the necessary HPC skills to find a suitable job by providing a central location

These three objectives have been addressed as described below.

9.3.1
Supporting Talent Generation

The **HPC Careers Case Studies**[99] feature the personal stories of people from different backgrounds working in a variety of HPC-related jobs. The case studies aim to raise general awareness of HPC as a skill which can lead to interesting jobs, and to break down any preconceived ideas about the types of people who work in HPC. By using first-hand accounts they also aim to highlight what people find exciting and rewarding about working in this field.

The **Report on Promotion of HPC as a Career Choice** addressed two different aspects of ensuring that there are enough people in the workforce with HPC skills: ***attracting*** young talent to careers in HPC, and ***retaining*** skilled workers in the HPC sector.

· **To attract new talent**, the following areas should be prioritised:

> **Public engagement**: raising the profile of HPC in general;
> **Promoting HPC career opportunities**: raising the profile of the variety of HPC-related jobs that exist and encouraging activities such as Women in HPC[100] and Diversity in HPC[101] which reach out to groups of people that are under-represented in the sector;
> **Better promoting the available opportunities**: e.g. by providing a centralised place to advertise job vacancies (such as the EXDCI Job Portal);
> **Building a community**: through Champions-style initiatives, such as ARCHER Champions[102], XSEDE Campus Champions[103], and the EDISON Education and Training Champions [104];
> **Integrating HPC into undergraduate and postgraduate** courses to expose all students in relevant discipline to basic HPC skills;
> Increasing alternative opportunities for learning, e.g. summer schools, online learning, etc;
> Promoting general computing as an essential skill to university and school-age students, to give everyone the foundations on which to build HPC expertise.

· **To retain existing talent** within the HPC workforce, there needs to be some recognition of the role, and a clear career path needs to be established. The report looks at two examples where this has been done in similar fields: the establishment of the Research Software Engineer (RSE), driven by the Software Sustainability Institute[105], and the work by the EDISON Data Science Framework[106] to establish the Data Scientist as a profession. Putting into place an equivalent career framework for HPC professionals could help both to recruit and to retain talented people within the HPC workforce.

[98] https://exdci.eu
[99] https://exdci.eu/jobs-and-training/hpc-career-case-studies
[100] http://www.womeninhpc.org
[101] http://www.hpc-diversity.ac.uk
[102] http://www.archer.ac.uk/community/champions
[103] https://www.xsede.org/campus-champions
[104] http://edison-project.eu/edison/education-training-champions
[105] https://www.software.ac.uk/
[106] http://edison-project.eu/

9.3.2
Identifying and Meeting Future Training Needs

The EXDCI **Training Portal**[107] is a searchable portal containing announcements about upcoming training opportunities around Europe. It also contains a list of links to other course providers and training opportunities, including short courses, summer schools, webinars, MOOCs, archived video tutorials, sets of course slides, and other documentation.

The **Training Roadmap** identified four challenges in terms of being able to provide enough appropriate and accessible training in order to equip the future workforce with the necessary HPC skills. It also provided a list of Recommended Actions to address these challenges. The challenges identified are:

- Coping with rising demand;
- Addressing the widening target audience;
- Integrating HPC training into undergraduate / postgraduate curricula;
- Encouraging interdisciplinary working.

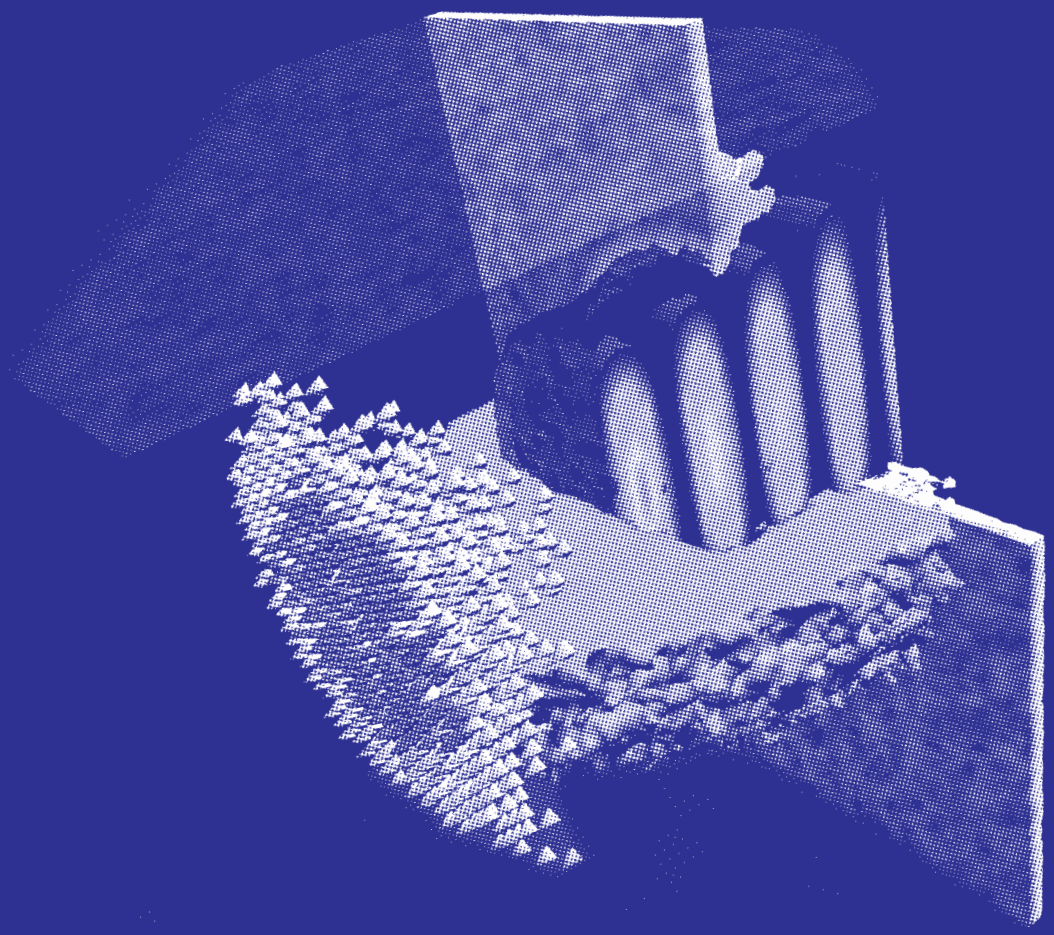
9.3.3
Facilitating HPC Staff Recruitment

The **EXDCI Job Portal**[108] is a searchable job portal which serves as a central point where job-seekers can search for suitable job vacancies, and employers can post their job advertisements in order to reach a wider audience. The Job Portal also contains a list of links to other places where job vacancies in HPC and related fields are posted, and some guidance for employers, to help them prepare their job descriptions.

[107] https://exdci.eu/jobs-and-training/training-portal

[108] https://exdci.eu/jobs-and-training/job-portal

10.



CONCLUSIONS AND OUTLOOK

This document provides a research roadmap for the remaining years of the H2020 framework programme. The awareness for the key role HPC plays in science, industry and in our everyday lives has increased dramatically of the past five years. The investments made on a pan-European public level demonstrate that HPC is not anymore a narrow niche of pure technical computing.

This effort of road-mapping Europe's HPC technology strategy needs to continue, despite the existence of other tools and mechanisms aimed at producing competitive European supercomputers. ETP4HPC welcomes and supports these initiatives but also claims that there will be a persistent need for basic research, prototyping and aligning priorities with other stakeholders, all of which should take place in the context of European competitiveness in science and industry. This is one of the roles the future SRAs should perform.

It is obvious that High Performance Computing, the vast number of Big Data use cases and fast-growing world of the "Internet of Things" cannot be seen as three separate silos. While each of the domains will keep having is own focus areas and priorities in the future, the interdependence is explicit in the case of important large projects such as SKA, autonomous driving, energy management, etc. In this context embedded and networking technologies will play a considerable role.

Therefore, on a European level, the forthcoming road-mapping efforts also need to be carried out in much closer cooperation between private/public bodies covering the domains of HPC, Big Data and IoT including expert groups representing the embedded and networking domains. The work performed lately together with BDVA and HiPEAC is a start into this direction.

11.

REFERENCES

EC 1 — Societal Challenges, https://ec.europa.eu/programmes/horizon2020/en/h2020-section/societal-challenges

EC 2 — Van der Pyl, Thierry, *The European HPC Strategy and actions in Horizon 2020*, presentation shown in Pisa on 5 Feb 2014.

EC 3 — High-Performance Computing: Europe's place in a Global Race, Brussels, 15.2.2012, http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2012:0045:FIN:EN:PDF

EC 4 — Communication from the commission to the European Parliament, the Council, the European economic and social committee and the committee of the regions, *European Cloud Initiative - Building a competitive data and knowledge economy in Europe*, Brussels, 19.4.2016

EC 5 — https://ec.europa.eu/digital-single-market/en/news/europe-global-player-high-performance-computing

EC 6 — *Implementation of the Action Plan for the European High-Performance Computing Strategy*, Accompanying the document: Communication from the Commission to the European Parliament, The Council, The European Economic and Social Committee and The Committee of the Regions, European Cloud Initiative — Building a competitive data and knowledge economy in Europe, {COM(2016) 178}

EC 7 — https://ec.europa.eu/commission/commissioners/2014-2019/oettinger/blog/luxembourg-launches-supercomputing-project_en

EC 8 — https://ec.europa.eu/digital-single-market/en/news/eurohpc-initiative-speeds-its-pace

EC 9 — https://ec.europa.eu/programmes/horizon2020/en/news/eight-new-centres-excellence-computing-applications

Ezell S.J and Atkinsion R.D, *The Vital Importance of High-Performance Computing to U.S. Competitiveness*, Innvation Technology & Innovation Institute, April 2016.

Hyperion — *Trends in the Worldwide HPC Market*, presented at ISC17, June 2017

HiPEAC — *HiPEAC Vision Document 2017*, https://www.HiPEAC.net/v17

HPC User Forum, *IDC Economic Models Linking HPC and ROI*, http://www.hpcuserforum.com/ROI/

IDC 1 — *High Performance Computing in the EU: Progress on the Implementation of the European HPC Strategy*, Final Report, Contract number: 30-CE-0663100/00-22, SMART number: 2014/0021

IDC 2 — Earl Joseph, Steve Conway, and Robert Sorensen, *IDC HPC ROI Research Update: Economic Models For Financial ROI and Innovation From HPC Investments*, (IDC, August 2015), https://hpcuserforum.com/presentations/colorado-sept2015/PublicIDCDOEROIResearchUpdate8.19.2015.pdf

IDC 3 — Earl C. Joseph, Chirag Dekate, and Steve Conway, *Real-World Examples of Supercomputers Used for Economic and Societal Benefits: A Prelude to What the Exascale Era Can Provide, (*IDC, May 2014), http://casc.org/wp- content/uploads/2014/07/IDCReportRealWorldExamplesOfBenefitsOfSupercomputers.pdf.

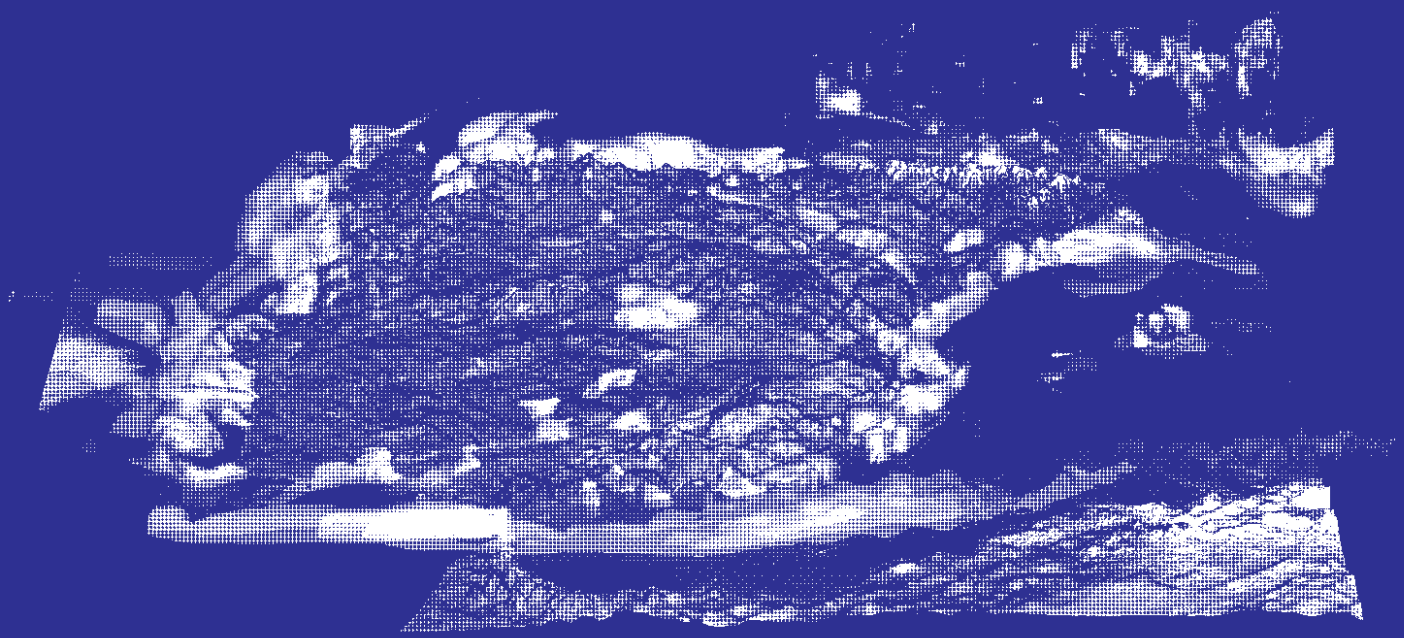
IDC 4 — *Major Trends in the Worldwide HPC Market*, Presentation at HPC User Forum at HLRS, February28 -March 1, 2017, Earl Joseph and Steve Conway

PRACE — PRACE Scientific Case for HPC in Europe 2012 – 2020, http://www.prace-ri.eu/prace-the-scientific-case-for-hpc/

Reed, D. A. & Dongarra, J. - Exascale Computing and Big Data Commun. ACM, ACM, 2015, 58, 56-68.

Statista, Europäische Union: Anteile der Wirtschaftssektoren am Bruttoinlandsprodukt (BIP) von 2005 bis 2015, https://de.statista.com/statistik/daten/studie/249078/umfrage/anteile-der-wirtschaftssektoren-am-bruttoinlandsprodukt-bip-der-eu/

12.



APPENDIX

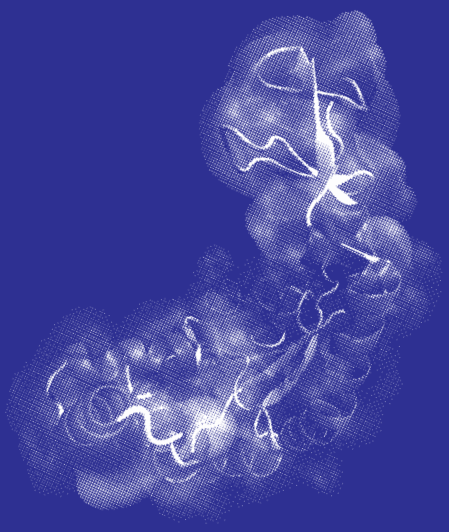
ETP4HPC SWOT (Strengths, Weaknesses, Opportunities, Threats) Analysis (May 2017)

The following SWOT analysis on the European HPC Technology Value Chain (with an objective to increase its global market share) summarises the strategic positions of European HPC Technology. It was carried out by the ETP4HPC Steering Board in May 2017. It is not exhaustive and it does shape the contents of this SRA directly. However, it expresses the view of ETP4HPC in certain areas.

STRENGTHS	OPPORTUNITIES
Leading industrial HPC users in applications and software e.g. Weather Modelling, Material Sciences, etc.	Well-managed programme(s) focused on societal impact
Focus on the advancement of science & technology, not Flops	Invest in completing the HPC value chain (where we are weak)
Political commitment	Focus on emerging domains such as precision medicine and gene sequencing
System integrator ecosystem	Creating machines to advance specific scientific & technical domains (linking HPC to the users)
Diverse and strong and HPC-aware industry users	Opportunity to address untapped markets – etc. Africa (however, no industrial ecosystem exists there). China and others might move faster

WEAKNESSES	THREATS
EU is collection of separate countries with individual strategies	Applications not ready for New infrastructure innovations (Heterogeneity, NVRAM, etc.)
Lack of HPC education and engaging with academia,	Companies completely relying on public funding (startups especially) – public funding should rather be a catalyst
Not enough encouragement for startups and SMEs in HPC	European hardware Technology (e.g. processors) experience higher risk to move further. Ecosystem not convinced about their value. Industry relies heavy on non-European IP.
Industrial ecosystem does not have complete coverage (for a full HPC stack needed for extreme scale)	Industry not being able to find the proper components in the value chain
Lack of private funding for HPC - funding for ecosystem development	No policy for addressing other untapped markets – etc., e.g. Africa (however, no industrial ecosystem exists there). China and others might move faster
Prioritisation for the needs of Big Data might cause HPC to lose priority	No Amazon, Google, etc. – like services in Europe (or much smaller). They might take up the "base" of the pyramid (need to encourage people like OVH)
The CoEs are doing a tremendous job – they have become a credible partner representing the European HPC Application expertise. However, their effort lacks long-term continuity, resulting in some good work being lost in between projects.	
While the current (and potential) HPC technology projects cover most important areas, due to the competitive and open project selection process, some areas are not covered or not given enough attention. Also, the entire effort suffers from fragmentation and duplication. A centralised project control mechanism is needed.	

13.



CONTRIBUTORS

The SRA is the work of many experts who took part in workshops, conference calls and other interactions. We would like to thank them for their contributions. Below, we only list the persons responsible for various parts of the document.

ETP4HPC Chairman
Jean-Pierre Panziera

ETP4HPC SRA 3 Editorial Team
Michael Malms (Coordinator)
Jean-Philippe Nominé
Maike Gilliot
Marcin Ostasz

SRA 3 Working Group Leaders

Balance, Compute I/O & Storage Performance	Sai Narasimhamurthy
Big Data & HPC Use Models	Costas Bekas
Energy & Resilience	Igor Zacharov
Extreme-Scale Demonstrators	Marc Duranton, Thomas Eickermann
HPC System Architecture & Components	Laurent Cargemel
Mathematics & Algorithms	Dirk Pleiter
Programming Environment	Paul Carpenter, Guy Lonsdale
Systems Software & Management	Hans-Christian Hoppe, Pascale Rosse-Laurent

Other Contributors
Alfred Geiger
Catherine Inglis
David Henty
Erwin Laure
François Bodin
Frank van der Hout
Jean-François Lavignon
Mark Asch
Stephane Requena

IMPRINT

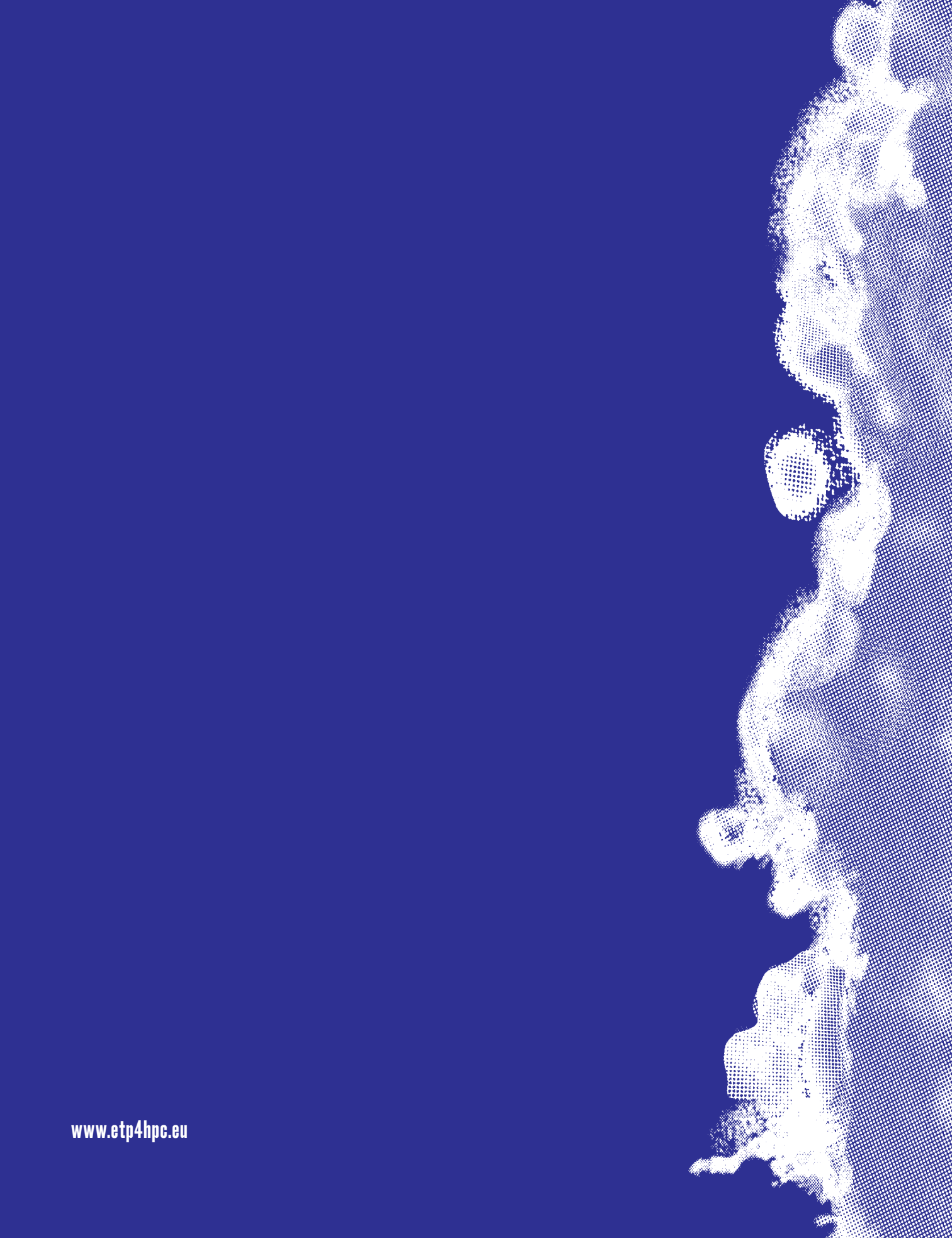
Text:
ETP4HPC

Graphic design and layout:
WWW.WORKSHIP.ES

Paper:
BIO TOP 250 G/M2
BIO TOP 120 G/M2

Printed and bound in Barcelona in February 2018:
WWW.AGPOGRAF.CAT

Contact ETP4HPC:
OFFICE@ETP4HPC.EU



www.etp4hpc.eu