DEEP

Extreme Scale Technologies

H2020-FETHPC-01-2016

DEEP-EST

DEEP - Extreme Scale Technologies

Grant Agreement Number: 754304

D7.4

Policy Briefing

Final

Version:	1.0
Author(s):	E. Gellner (BADW-LRZ)
Contributor(s):	
Date:	28.06.2019

Project and Deliverable Information Sheet

DEEP-EST Project	**Project Ref. №:** 754304	
	Project Title: DEEP - Extreme Scale Technologies	
	Project Web Site: http://www.deep-projects.eu	
	Deliverable ID: D7.4	
	Deliverable Nature: Other	
	Deliverable Level: PU *	**Contractual Date of Delivery**: 30 / 06 / 2019
		Actual Date of Delivery: 28 / 06 / 2019
	EC Project Officer: Juan Pelegrín	

* - The dissemination levels are indicated as follows: PU = Public, fully open, e.g. web; CO = Confidential, restricted under conditions set out in Model Grant Agreement; CI = Classified, information as referred to in Commission Decision 2001/844/EC.

Document Control Sheet

Document	**Title:** Policy Briefing	
	ID: D7.4	
	Version: 1.0	**Status:** Final
	Available at: http://www.deep-projects.eu	
	Software Tool: Microsoft Word	
	File(s): DEEP-EST_D7.4_PolicyBriefing_v1.0	
Authorship	**Written by:**	E. Gellner (BADW-LRZ),
	Contributors:	
	Reviewed by:	E. Suarez (JUELICH), Julita Corbalan (BSC)
	Approved by:	BoP/PMT

Document Status Sheet

Version	**Date**	**Status**	**Comments**
1.0	28/06/2019	Final version	EC submission

Document Keywords

Keywords:	DEEP-EST, HPC, Exascale, Policy briefing, Dissemination

Table of Contents

Project and Deliverable Information Sheet 1

Document Control Sheet 1

Document Status Sheet 2

Document Keywords 3

Table of Contents 4

Executive Summary 5

1 Introduction: Benefits of HPC in Europe 5

2 The need of a Modular Supercomputing Architecture 6

3 DEEP Projects: Towards a Modular Supercomputing Architecture 6

3.1 DEEP-EST ambition 6

3.2 MSA and its innovation potential 7

3.3 Current project status and expected impacts 8

4 Conclusion 9

List of Acronyms and Abbreviations 11

Executive Summary

The deliverable D7.4 represents a public policy briefing for political stakeholders and lobbying organisations detailing conclusions from project results and impact on High Performance Computing (HPC) development in Europe.

First, challenges in Europe will be discussed in an introduction paragraph and why HPC is needed to improve the situation in the scientific, societal and economic sector. The aim of DEEP-EST is to create a Modular Supercomputing Architecture (MSA) to address the needs that will be described. Before the current status and the impact of the DEEP-EST project on the development of HPC in Europe will be explained, the advantages of MSA will be discussed in section two.

1 Introduction: Benefits of HPC in Europe

Societal, scientific and economic needs are the drivers for the next generation HPC with exascale performance.

Industry and SMEs are increasingly relying on the power of supercomputers to invent innovative solutions, reduce cost and decrease time to market for products and services. The European car industry for example provides jobs for twelve million people and accounts for four percent of the EU gross domestic product. European car manufacturers will need more computing capacity in the future for autonomous driving and digitization.

Furthermore, all scientific disciplines are becoming computational today. To answer fundamental science questions, make new discoveries and breakthroughs, very high computing power and capability to deal with huge volumes of data is needed. HPC and Big Data analysis provide scientists with deeper insights into unexplored areas and systems of the highest complexity.

Last but not least, HPC brings insight into complex topics, contributing to make them generally understandable and thus bringing them closer to society. For example simulations to reduce the environmental footprint and predicting the impact of severe weather conditions.

HPC is part of a global race. Many countries have announced ambitious plans for building the next generation HPC with exascale performance and deploying state-of-the-art supercomputers. The available supply of computation time cannot satisfy an ever growing demand. To fill the gap, European scientists and industry increasingly process their data outside the EU. This can create problems related to privacy, data protection, commercial trade secrets, and ownership of data in particular for sensitive applications. Europe consumes about 29 percent of HPC resources worldwide today, but the EU industry provides only ~5 percent of such resources. But security is not the only important factor. Independence is strategically important in order to pursue Europe's commercial interests and visions.

HPC is a strategic resource for Europe's future as it enables researchers to study and understand complex phenomena while allowing policy makers to make better decisions and enabling industry to innovate in products and services. The EU response to the above is to invest together in an ambitious supercomputing infrastructure strategy. The EU's ambition is to become one of the world leaders in supercomputing. The European Commission funds projects to address these needs. One of this EU funded projects is DEEP-EST that creates a modular supercomputing architecture.[1]

2 The need of a Modular Supercomputing Architecture

Scientists and engineers run large simulations on supercomputers to describe and understand problems too complex to be reproduced experimentally. The codes that they use for this purpose, the kind of data they generate and analyse, and the algorithms they employ are very diverse. As a consequence, some applications run better (faster, more cost- and more energy-efficient) on certain supercomputers and some run better on others. The better the hardware fits the applications (and vice-versa), the more results can be achieved in the lifetime of a supercomputer. But finding the best match between hardware technology and the application portfolio of HPC centres is getting harder. Furthermore, some supercomputing centres have different, seperated clusters and port their applications to smaller clusters as for example at the Supercomputing Centre in Barcelona. But the fact that the different clusters are seperated is limiting the potential of that applications.

Computational science and engineering keep advancing and address ever-more complex problems. To solve these problems, research teams frequently combine multiple algorithms, or even completely different codes, that reproduce different aspects of the given topic. Furthermore, new user communities of HPC systems are emerging, bringing new requirements. This is the case for large-scale data analytics or big data applications: They require huge amounts of computing power to process the data deluge they are dealing with. Both complex HPC workflows and HPDA applications increase the variety of requirements that need to be properly addressed by a supercomputer centre when choosing its production systems. These challenges add to additional constraints related to the total cost of the machine, its power consumption, the maintenance and operational efforts, and the programmability of the system.

The Modular Supercomputer Architecture is an innovate approach developed in Europe to build High-Performance Computing systems by coupling various compute modules, following a building-block principle. Each module is tailored to the needs of a specific group (or parts) of applications, and all modules together behave as a single machine. This is ensured by connecting them through a high-speed network and operating them with a uniform system software and programming environment – its Network Federation. This allows one application to be distributed over several modules, running each part of its code onto the best suited hardware module. With such an adaptable system one can build computers fulfilling the needs of a broad range of users, and at the same time achieve "exascale" performance – one billion billion (i.e. a quintillion) operations per second.

3 DEEP Projects: Towards a Modular Supercomputing Architecture

3.1 DEEP-EST ambition

The DEEP-EST project will implement the Modular Supercomputer Architecture, and it will conduct a thorough evaluation of such a system with six ambitious applications that combine HPC simulation and large scale HPDA. DEEP-EST aims to improve the usability, flexibility,

[1] https://eurohpc-ju.europa.eu/index.html#inline-nav-4

and sustained performance of future supercomputers significantly beyond what can be provided by today's monolithic systems. Across several technology fields, significant progress beyond the state of the art is required in order to attain the flexibility to combine modules optimised for specific application areas:

- Data Analytics Module incorporating cutting-edge acceleration and non-volatile memory technology and designed to best support large scale, HPDA applications.
- Combination of modules with different interconnects: the project will demonstrate the Network Federation between them. A highly optimised bridging protocol between Mellanox InfiniBand and EXTOLL will be developed, significantly extending the results from the DEEP project.
- Efficient management of heterogeneous resources. The ParaStation Resource Management will be extended to handle arbitrary combinations of resources across the modules.
- Efficient co-scheduling of resources spread across different modules: The project will extend the SLURM scheduler to allow for a flexible and scalable operation of a MSA system for both, applications utilising different modules at the same time and work-flow oriented approaches. The utilization of SLURM and other system services and/or programming models widely used in current systems will help to simplify the adoption of the DEEP-EST contributions.
- Cluster Module speeding-up the less scalable applications (or parts thereof) with need for high single-thread performance, while the large scaling and data-intensive parts run on the ESB and DAM, respectively.
- Data Analytics frameworks and programming models that fully leverage the clustered DAM and support scale-out of large data analytics problems, integrated with HPC programming models, where it makes sense.
- The ESB is expected to become the largest module in the DEEP-EST prototype and will be tailored to the needs of highly scalable (parts of) applications. Because of that, special focus will be put into realising a highly scalable system from the hardware and software point of view, with energy efficiency as key aspect of the system integration.

3.2 MSA and its innovation potential

The MSA is a novel, European approach to overcome the fundamental limitations of today's monolithic supercomputer architectures with respect to flexibly and efficiently serving: heterogeneous applications like multi-physics simulations, combinations of classical HPC simulations with HPDA, and the large mix of applications with different resource requirements typically run by a supercomputer centre. The architecture combines innovative hardware and software elements. The potential of this innovation is immense: if successful, throughput for a centre would increase and TCO and energy required for the above-mentioned applications would be substantially reduced. Looking further into the future, the results would greatly influence the way of harnessing von Neumann architectures and make them accessible to users in science and industry. In addition, the component modules and the Network Federation will drive innovation and achieve scalability and efficiency gains over the best of breed systems today. Of particular interest are the advances in interconnect technology (EXTOLL) and the Network Federation, the acceleration of collective communication (GCE), the provision of globally accessible memory resources (NAM), the novel clustered Data Analytics Module (DAM) itself and the Extreme Scale Booster (ESB) for the highly- scalable parts of the applications. These innovations will ensure that HPC centres can make full use of their installed system capabilities and capacity.

3.3 Current project status and expected impacts

The project technology results will directly contribute to the implementation of the Strategic Research Agenda laid down by the ETP4HPC. At this point of the project execution, it is not yet possible to quantify the impact of the results created so far, since system manufacturing and installation have not yet been completed – Although the recent installation of the first module (Cluster Module) at the Supercomputing Center Jülich has brought DEEP-EST one step closer to its goal.[2] Key project partners (e.g. JUELICH, BSC, Intel, Megware, FHG-ITWM, ParTec, etc.) are deeply involved in writing and updating the SRA, the third edition of which became available in 2017. This strong link between the ETP4HPC and DEEP-EST ensures that the project results will fully align with the SRA priorities, remain relevant for the Exascale era and have a positive impact on the European HPC ecosystem.

Furthermore, a subset of DEEP-EST partners is also actively involved in initiatives addressing the goals set by the EuroHPC Joint Undertaking (JU), which promotes the development of European technologies for pre-Exascale and Exascale systems. The DEEP-EST project in itself aligns perfectly to this strategy, by advancing development and validation of key European technologies. Examples are

- the EXTOLL network (with the Fabri[3] integrated fabric switch and the associated memory technologies GCE and NAM);
- the energy efficient integration of components from the integrator Megware;
- the cluster management and programming tools from ParTec;
- the file system BeeGFS by FHG-ITWM (with commercial support from its spin-off ThinkParQ);
- the OmpSs programming environment and the Extrae/Paraver/Dimemas performance analysis and modelling tools by BSC;
- the JUBE benchmarking environment and SIONlib I/O concentrator library by JUELICH.

The ESB will demonstrate an innovative way of building a highly scalable and efficient Booster out of GPGPU accelerators, general-purpose CPUs and leading European network technology. It is worth mentioning that, since the European Processor Initiative (EPI) is working on a CPU plus accelerator architecture, it will be able to profit from the DEEP-EST results, discoveries and innovations. Porting and optimisation of the pilot applications for the ESB will give these a head-start to run and scale with high efficiency on future EPI-based systems. Therefore, DEEP-EST is fully aligned with the EuroHPC objectives.

The DEEP-EST concept and its constituent technologies will be demonstrated and validated by the implementation of a system prototype with a complete SW environment and by the use of six important European workloads, which combine HPC-style computation with data analytics and machine learning. Following a strict co-design approach, these six applications did shape the system architecture and design (for both HW and SW), and they do represent a variety of important research areas expected in future workloads for HPC centres. As the integrated prototype system becomes available, the six workloads will be adapted and

[2] https://www.deep-projects.eu/press-corner/news/317-press-release-towards-flexible-exascale-computing-installation-of-the-first-deep-est-module-by-megware.html

optimised, and they will serve as the touchstone for validating the DEEP-EST architecture promise.

The stringent co-design approach clearly is an important legacy of the DEEP and DEEP-ER projects. It is a proven scheme of jointly developing the HW, system SW, and applications and ensuring that the resulting prototype system will be well integrated and best support the evolved applications. The successful application of this method creates a precedent and example that will impact future HPC projects.

The other tried and proven approach from DEEP and DEEP-ER taken up by DEEP-EST is the provision of early development systems and evaluators that are representative of the prototype modules. A first list of such systems has been compiled. Some have been used already to evaluate processor and network technologies considered candidate-components of the DEEP-EST prototype. Application use cases where used to run different benchmarks. The outcome of these tests has helped deciding the final configuration.

The DEEP-EST architecture addresses the needs of medium to large-scale supercomputer centres, which all run a large variety of applications and which see a trend towards combining HPC computations (such as simulation) with large-scale data analytics and machine learning. The technology implemented for the DEEP-EST prototype will be directly applicable, and it will bring attractive performance and efficiency improvements compared to the state of the art in homogeneous systems or systems that use heterogeneity only within Cluster Nodes.

The DEEP-EST technology results will also be relevant for the “Departmental/Divisional” class of HPC systems as they are vital for European SMEs and corporate enterprises in the growing digital economy. In fact, partner Megware is one of the most commercially successful European providers of cluster computers and technology for Universities, and small-, medium, and large-scale academic and commercial computer centres. This role will facilitate the introduction of DEEP-EST technologies into a wider market.

The DEEP-EST architecture aims to satisfy the needs of both HPC and large-scale data analytics/machine learning applications and provide a perfect execution system for combined applications and diverse workload mixes. The DEEP-EST project is building an integrated prototype system, tailored to best support the co-design applications, which in turn are representative of emerging, innovative HPC and HPDA codes. The project results will therefore enable users in academia and industry to reap the benefits of combining proven HPC computation with cutting-edge data analytics and machine learning to help create breakthrough scientific discoveries and engineering results. In this context, DEEP-EST has a large potential to produce a strong impact in the academic and industrial HPC community, targeting both end-users and computer centres.

4 Conclusion

Making the heterogeneity homogeneous – The optimization of homogeneous HPC systems has more or less reached its limit. In addition to compute-intensive simulations and the traditional tasks undertaken in scientific computing centers, new applications such as big data analytics and sophisticated visualizations are gaining importance – but current supercomputer architectures cannot handle these tasks efficiently and at the same time continue serving the large-scale simulations that constitute their traditional workloads.

This report does not only explain the importance of high performance computing in general for industry, research and society. Rather, it shows why the Modular Supercomputing Architecture in particular can deliver the best possible results for a diverse application portfolio, and how the DEEP-EST project is implementing the MSA. The third member of the DEEP Projects family builds upon the results of its predecessors DEEP and DEEP-ER, which ran from December 2011 to March 2017. An aim uniting all three projects from the beginning: to gradually develop the prerequisites for the highly efficient Modular Supercomputing Architecture that can be flexibly adapted to the various requirements of scientific applications with highest efficiency and scalability.

List of Acronyms and Abbreviations

B

BADW-LRZ:	Leibniz-Rechenzentrum der Bayerischen Akademie der Wissenschaften. Computing Centre, Garching, Germany
BeeGFS:	The Fraunhofer Parallel Cluster File System (previously acronym FhGFS). A high-performance parallel file system.
BoP:	Board of Partners for the DEEP-EST project

C

CM:	Cluster Module: with its Cluster Nodes (CN) containing high-end general-purpose processors and a relatively large amount of memory per core
CN:	Cluster Node (functional entity)
CPU:	Central Processing Unit

D

D:	Deliverable, followed by a number, term to designate a deliverable (document) in the DEEP-EST project
DAM:	Data Analytics Module: with nodes (DN) based on general-purpose processors, a huge amount of (non-volatile) memory per core, and support for the specific requirements of data-intensive application
DEEP:	Dynamical Exascale Entry Platform (project FP7-ICT-287530)
DEEP-ER:	DEEP - Extended Reach (project FP7-ICT-610476)
DEEP/-ER:	Term used to refer jointly to the DEEP and DEEP-ER projects
DEEP-EST:	DEEP - Extreme Scale Technologies
Dimemas:	Performance analysis tool developed by BSC
DN:	Nodes of the DAM

E

EC:	European Commission
ETP4HPC:	European Technology Platform for High Performance Computing
ESB:	Extreme Scale Booster: with highly energy-efficient many-core processors as Booster Nodes (BN), but a reduced amount of memory per core at high bandwidth
EU:	European Union

Exascale:	Computer systems or Applications, which are able to run with a performance above 10^{18} Floating point operations per second
EXTOLL:	High speed interconnect technology for HPC developed by UHEI
Extrae:	Performance analysis tool developed by BSC

F

fabri^3:	Interconnect technology based on EXTOLL (pron. “Fabri-Cube”)
FHG-ITWM:	Fraunhofer Gesellschaft zur Foerderung der Angewandten Forschungs e.V., Germany

G

GCE:	Global Collective Engine, a computing device for collective operations
GPGPU:	General Purpose Graphics Processing Unit

H

H2020:	Horizon 2020
HBM:	High Bandwidth Memory
HPC:	High Performance Computing
HPDA:	High Performance Data Analytics
HW:	Hardware

I

InfiniBand:	A networking communication standard for HPC clusters
Intel:	Intel Germany GmbH, Feldkirchen, Germany
I/O:	Input/Output. May describe the respective logical function of a computer system or a certain physical instantiation

J

JUBE:	Jülich Benchmarking Environment
JUELICH:	Forschungszentrum Jülich GmbH, Jülich, Germany

M

M:	Month, followed by a number, term to designate a duration month in the DEEP-EST project (relative to the stating date)
MB:	Mega Bytes

Megware:	Megware Computer Vertrieb und Service GmbH, Chemnitz, Germany
MPI:	Message Passing Interface, API specification typically used in parallel programs that allows processes to communicate with one another by sending and receiving messages
MSA:	Modular Supercomputer Architecture

N

NAM:	Network Attached Memory
NIC:	Network Interface Controller
NVM:	Non-Volatile Memory. Used to describe a physical technology or the use

O

OmpSs:	BSC's Superscalar (Ss) for OpenMP
OpenMP:	Open Multi-Processing, Application programming interface that support multiplatform shared memory multiprocessing

P

ParaStation:	Software for cluster management and control developed by JUELICH and its linked third party ParTec
Paraver:	Performance analysis tool developed by BSC
ParTec:	ParTec Cluster Competence Center GmbH, Munich, Germany. Linked third Party of JUELICH in DEEP-EST
PI:	Principal Investigator
PMT:	Project Management Team of the DEEP-EST project

S

SIONlib:	Parallel I/O library developed by Forschungszentrum Jülich
SLURM:	Job scheduler that will be used and extended in the DEEP-EST prototype
SME:	Small and Medium Enterprises
SRA:	Strategic Research Agenda prepared by ETP4HPC
SW:	Software

T

ThinkParQ:	Spin-off company of FHG-ITWM

U

UHEI:	Ruprecht-Karls-Universitaet Heidelberg, Germany

W

WP:	Work package