H2020-FETHPC-3-2017 - Exascale HPC ecosystem development

EXDCI-2

European eXtreme Data and Computing Initiative - 2

Grant Agreement Number: 800957

D2.3
Big Data, embedded and edge computing and HPC synergies
Final

Version: 1.0
Author: Michael Malms, ETP4HPC
Date: 18/08/2020

Project and Deliverable Information Sheet

EXDCI Project	**Project Ref. №: FETHPC-800957**	
	Project Title: European eXtreme Data and Computing Initiative - 2	
	Project Web Site: http://www.exdci.eu	
	Deliverable ID: D2.3	
	Deliverable Nature: Report	
	Dissemination Level: PU (Public) *	**Contractual Date of Delivery:** 31/08/2020
		Actual Date of Delivery: 28/08/2020
	EC Project Officer: Evangelia Markidou	

* Public as referred to in Commission Decision 2991/844/EC

Document Control Sheet

Document	**Title: Big Data, embedded and edge computing and HPC synergies**	
	ID: D2.3	
	Version: 1.0	**Status:** Final
	Available at: http://www.exdci.eu	
	Software Tool: Microsoft Word 2016	
	File(s): D2.3-final.docx	
Authorship	**Written by:**	Michael Malms, ETP4HPC
	Contributors:	Marc Duranton, CEA Francois Bodin, IRISA Hans-Christian Hoppe, INTEL Gabriel Antoniu, INRIA Marc Asch, U-PICARDIE Zoltan Horvarth, EU-MATHS-IN Jens Krueger, FRAUNHOFER Peter Bauer, ECWMF
	Reviewed by:	Bernd Mohr, FZ JUELICH John Clifford, PRACE
	Approved by:	MB/TB

Document Status Sheet

Version	Date	Status	Comments
0.1	01/07/2020	Draft	First full text
0.2	21/07/2020	Draft	Various amendments
0.9	27/07/2020	Final Draft	Submission for review
1.0	18/08/2020	Final version	With fixes from reviews

Document Keywords

Keywords:	ETP4HPC, Strategic Research Agenda, Multi-Annual Roadmap, HPC Technologies, Research and Innovation, Digital Continuum, Transcontinuum, TransContinuum Initiative

Table of Contents

List of Figures

List of Tables

References and Applicable Documents

[1] https://www.exascale.org/bdec/
[2] Pathways to Convergence, IJHPCA, 32(4), 2018
[3] https://sylabs.io/docs/
[4] http://www.etp4hpc.eu/sra
[5] https://www.etp4hpc.eu/bigdata.html
[6] http://www.etp4hpc.eu
[7] www.bdva.eu
[8] https://ecs-org.eu/
[9] https://5g-ia.eu/
[10] https://claire-ai.org/
[11] https://www.eu-maths-in.eu/
[12] https://www.hipeac.net/
[13] https://cheese-coe.eu
[14] https://www.eocoe.eu
[15] https://www.dkrz.de/about-en
[16] https://www.excellerat.eu
[17] https://hidalgo-project.eu
[18] https://www.eu-maths-in.eu/wp-content/uploads/2018/05/MSO-vision.pdf

List of Acronyms and Abbreviations

AI	Artificial Intelligence
AIOTI	Alliance for the Internet of Things Innovation
BDEC	Big Data and Extreme-scale Computing
BDVA	Big Data Value Association
CoE	Centre of Excellence (for Computing Applications)
CPS	Cyber- Physical System
DoA	Description of Activity
DSL	Domain-Specific Languages
DT	Digital Twin
DX.Y	Deliverable Number X.Y
EC	European Commission
ECSO	European Cyber security Organisation
ETP4HPC	European Technology Platform for High Performance Computing
EU	European Union
EXDCI	European Extreme Data and Computing Initiative
FET	Future and Emerging Technologies
H2020	Horizon 2020 - The EC Research and Innovation Programme in Europe
HiPEAC	European Network of Excellence on High Performance and Embedded Architecture and Compilation
HPC	High Performance Computing
HPDA	High-Performance Data Analytics
HW	Hardware
I/O	Input/Output
IEEE	Institute of Electrical and Electronics Engineers
IoT	Internet of Things
IT	Information Technology
MFF	Multiannual Financial Framework
ML	Machine Learning
MOR	Model Order Reduction
MSODE	Modelling, Simulation and Optimisation in Data-rich Environment
QoS	Quality of Service
R&D	Research and Development
R&I	Research and Innovation
RIAG	EuroHPC R&I Advisory Group, one of the two Advisory Groups of EuroHPC Industrial and Scientific Advisory Board
SRA	Strategic Research Agenda
SRA4	Fourth edition of ETP4HPC's SRA
TCI	TransContinuum Initiative

Executive Summary

In the last decade, the High-Performance Computing landscape has undergone numerous changes and the role and definition of HPC itself has been 'a moving target' due to its interactions and inter-dependencies with other related areas. The complex task of integrating data-related features and functionalities (e.g. extreme data scale, sensors, HPDA, machine learning) as well as exascale computing into application workflows has required the stakeholders of the HPC Community to rethink how the cyberinfrastructure should be organised and used. A number of use cases identified within this task demonstrate how HPC and other technologies are interwoven within complex systems.

One of the first concepts that has shaped the definition of HPC-related research roadmaps has been the 'convergence' of the HPC and HPDA software stacks. Consequently, the use of container technologies has been identified as a solution to deal with the heterogeneity of the software tools needed for HPDA/AI. Furthermore, HPC technologies have been introduced to facilitate in situ data processing.

The notion of 'continuum' was introduced at a later stage to capture the fact that data are produced, processed and stored in many places (e.g. IoT, large scientific instruments in various infrastructures such as Clouds or at the Edge). A number of issues have been identified which are intrinsic to privacy and data locality optimisation. A further extension is the *'Transcontinuum'* concept, which encapsulates the fact that cyberinfrastructure elements cannot be used independently. It is necessary to provide a coherent and effective view of the cyberinfrastructure to end-to-end application developers. Achieving this vision requires addressing numerous challenges involving many communities.

For instance, an important challenge is to provide tools to deal with complex application workflows (addressing issues such as distributed programming, orchestration, transversal monitoring) to be deployed in a cyberinfrastructure environment which is altogether heterogeneous, multi-tenant, and multi-owner. In such a case, resource allocation and orchestration are critical issues (for instance, in the case of an application workflow that spans over an IoT network, a Cloud centre and a supercomputing centre).

Another essential challenge is data logistics optimisation. This topic includes managing data life cycle, storage, network, privacy properties, access control, metadata, as well as other issues. The global environmental challenge is strongly linked to this issue as a way to achieve sustainability, energy efficiency, resource saving, data reduction, algorithm efficiency, etc. Another consideration is the concept of 'AI everywhere', for instance, at the edge to improve data locality, bandwidth efficiency, privacy, but also across the entire continuum where AI is used for science from replacement or acceleration of computing models to data post-processing, as well as for cyberinfrastructure optimisation. Finally, to allow a safe deployment of applications on the cyberinfrastructure, cybersecurity needs to acquire a new dimension that encompasses transversal authentication, supervision, resilience, trusted communications, etc.

Most importantly, none of these challenges can be addressed without an extensive training support program.

A major next step in fostering a larger horizontal collaboration between the disciplines involved in the implementation of the Transcontinuum has been reached in July 2020. The "TransContinuum Initiative" (TCI) is being formed among seven leading European Associations (ETP4HPC, HiPEAC, BDVA, EU-MATHS-IN, 5G-IA, ECSO, CLAIRE). One of the main goals of this Initiative is to jointly prioritize R&I in support of the European Transcontinuum.

1 Introduction: HPC in an integrated, digital continuum

The rapid proliferation of digital data generators, the unprecedented growth in the volume and diversity of the data they generate, and the intense evolution of the methods for analysing and using that data are radically reshaping the landscape of scientific computing. The most critical problems involve the logistics of wide-area, multi-stage workflows that will move back and forth across the computing continuum, between the multitude of distributed sensors, instruments and other devices at the network's edge, and the centralised resources of commercial clouds and HPC centres [1]. The objective of this report is to put HPC into perspective of this new paradigm of 'The Digital Continuum'.

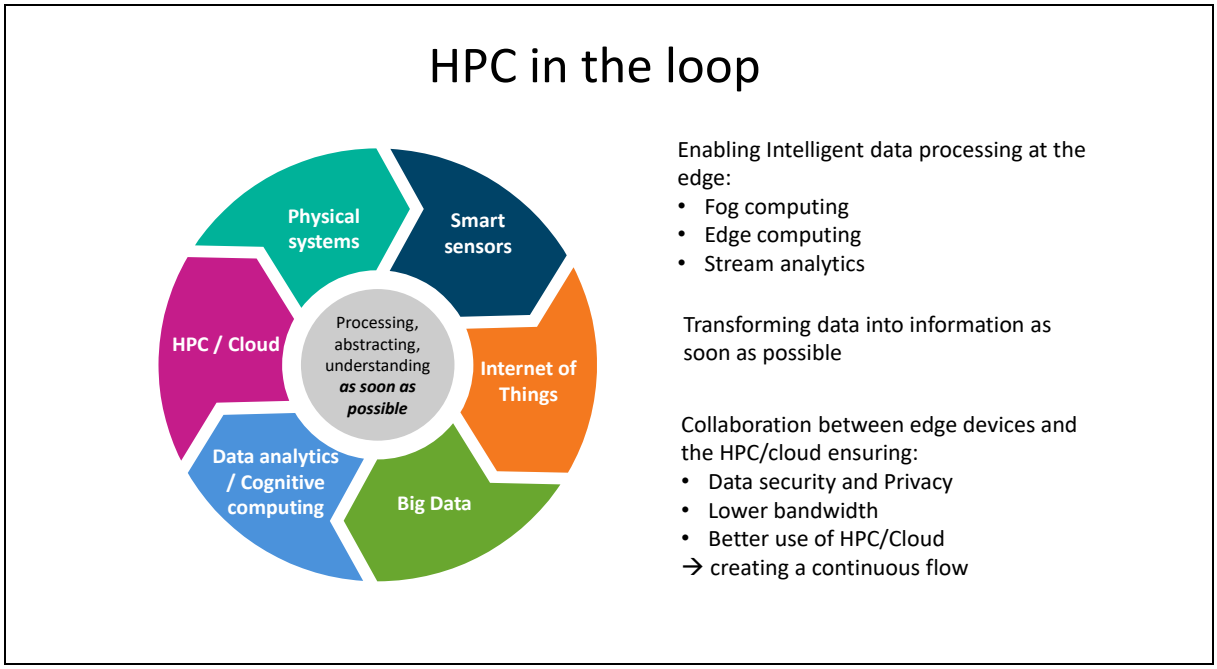

Figure 1: HPC in the loop (source HiPEAC)

1.1 Application and use case scenarios in a digital continuum

HPC has always been one of the main tools advancing science by delivering results, which can be attained by the use of cutting-edge computer technologies only. Throughout the last decade, numerical computing has been growing rapidly in many directions: higher fidelity, multi-physics models; deluge of observational data from sensors and of simulated data; semi-automatic data analysis and post-processing; uncertainty qualification and AI-based models. Combining all these aspects will result in a highly complex application (software) architecture, currently a focus area in related research.

In reference to Figure 1, this layer is driven by the thematic clusters and missions as well by industrial and scientific needs. The extraction of IT/HPC requirements out of representative and strategically important use case scenarios is necessary in order to drive HPC R&I in the right direction. They are key to assess new architectures or infrastructure as well as provide testbeds to research & industrial teams.

In the context of promoting innovations for the HPC, HPDA and IoT ecosystems, the use cases identified must be such that alignment with technology 'silos' is avoided, which would strongly

restrict the shaping capabilities for the R&I work program. Furthermore, fully addressing the societal challenge can only be achieved when considering end-to-end approaches where data production is integrated with data analytics, machine learning, numerical simulation, data archiving as well as the final use of the results. Underlying the use cases are applications relying on complex workflows within which individual tasks are executed on a wide variety of systems and whereby the complete data management cycle is addressed.

However, many representative use case scenarios are difficult to analyse since they combine many heterogeneous components (e.g. relying on different software stacks) as well as different resources or user governance strategies. For instance, this is about applications across a federation of systems - that includes HPC centres, cloud facilities, fog and edge components, and networks - while at the same time preserving security and privacy from end-to-end. Furthermore, the economics aspects of the deployment of these applications must be considered.

Consequently, this means facing extreme scale heterogeneity where, in the worst case, the common denominator may be a common governance and resource allocation policy. At a high level, the main technical challenges are how to achieve interoperability between the application workflow components, their orchestration, as well as reproducibility of execution in order to allow debugging and ease of deployment. In addition, infrastructure management and resource allocation policies are also strong roadblocks to overcome. For instance, supercomputers today are typically deployed in a way that they become silos, with limited external connectivity, proprietary access processes, relatively rigid operational models that expect users to submit batch jobs, and limited flexibility in terms of software stack provisioning. It is difficult to make them part of an application workflow that would include components deployed in the Cloud, handle streaming of data, for example.

1.2 Workflow and capabilities

Understanding the workflow and data flows is of crucial importance for an analysis of real use cases. Each use case (e.g. autonomous driving, personalised medicine, wind park operation, etc.) has its unique composition of basic 'functional capabilities':

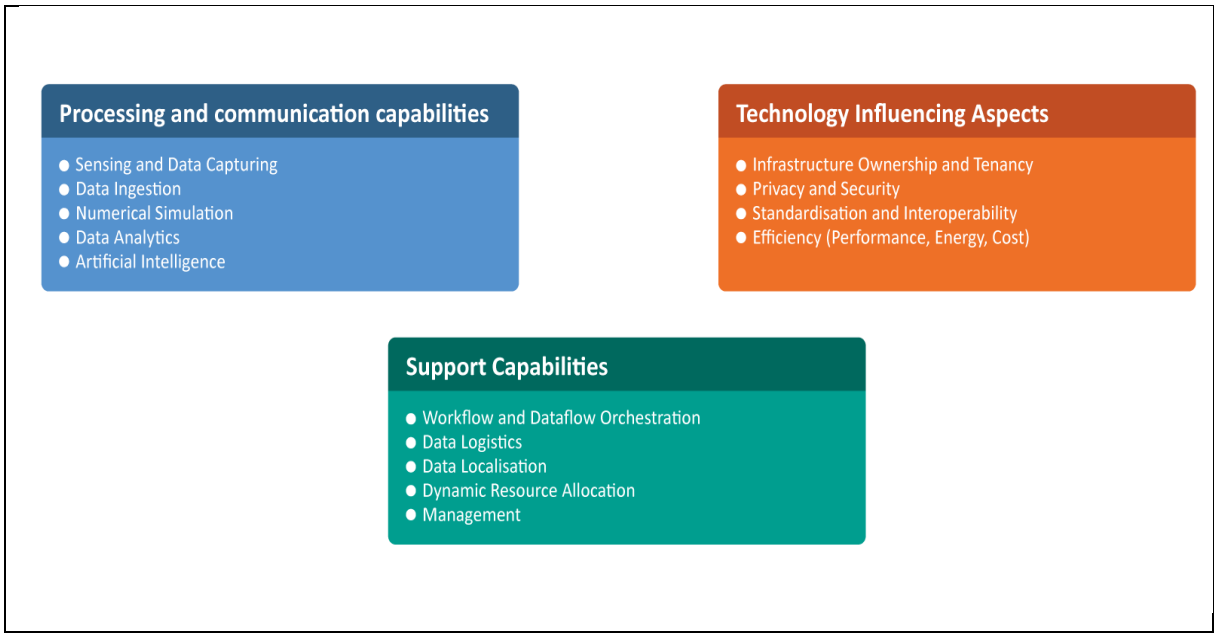

Figure 2: Categories of capabilities in mixed use scenarios

- The **'Processing and Communication Capabilities'** listed in Figure 2 cover all areas which require compute capabilities, be it in a datacentre, edge or fog node or an IoT device – each of them with a different application scope. For a given workflow (use case), the individual processing capabilities are expected to be spread meaningfully across locations and systems. We distinguish between data capture from devices, data ingestion into a compute environment, the typical HPC capability of numerical simulation, and the Big Data capabilities of data analysis and artificial intelligence. To address such new compute requirements, HPC capabilities must provide the processing capabilities for the Big Data environment, which includes interactive analytics as well as batch and real-time processing of data streams.
- The **'Technology Influencing Aspects'** are properties that significantly impact the design, implementation and integration of the processing capabilities but do not directly provide any data processing capabilities. They must be provided by the processing infrastructure in ways that satisfy the end-user requirements to guarantee an effective and efficient solution. The governance of compute infrastructure and data imposes policies on the data processing. Security and privacy must be considered in such an environment for most use cases to comply with regulatory and end-user needs. Interoperability and standards increase the trust in developed workflows and accelerate the adoption by users. The efficiency of a solution is relevant insofar that the costs of a solution limit the adoption in use-cases that yield limited revenue. A performant, energy and cost-efficient system maximizes industrial and commercial competence by enabling novel scenarios.
- **'Support Capabilities'** describe the crucial implementation aspects of a mixed scenario. As shown in Figure 2, the workflow reflects the interconnections of actions and data between the IoT devices, processing entities and data repositories. The identified capabilities are currently underdeveloped for the environment discussed here and require further R&D efforts.

The orchestration of workflows and automatic and efficient deployment across the complex hardware-landscape provided, is required to exploit such systems. For instance, data must be placed and migrated intelligently to match the storage and processing capabilities of (IoT or edge) systems. Finally, workflows must adapt their processing capabilities dynamically depending on the input, or other external parameters such as the number of users or availability of processing capacity. This requires software layers that enable such dynamic, ad-hoc changes. We recognise that management procedures must be developed that deal with the distributed nature of computation, ownership, and conformance to standards while considering the efficiency aspects.

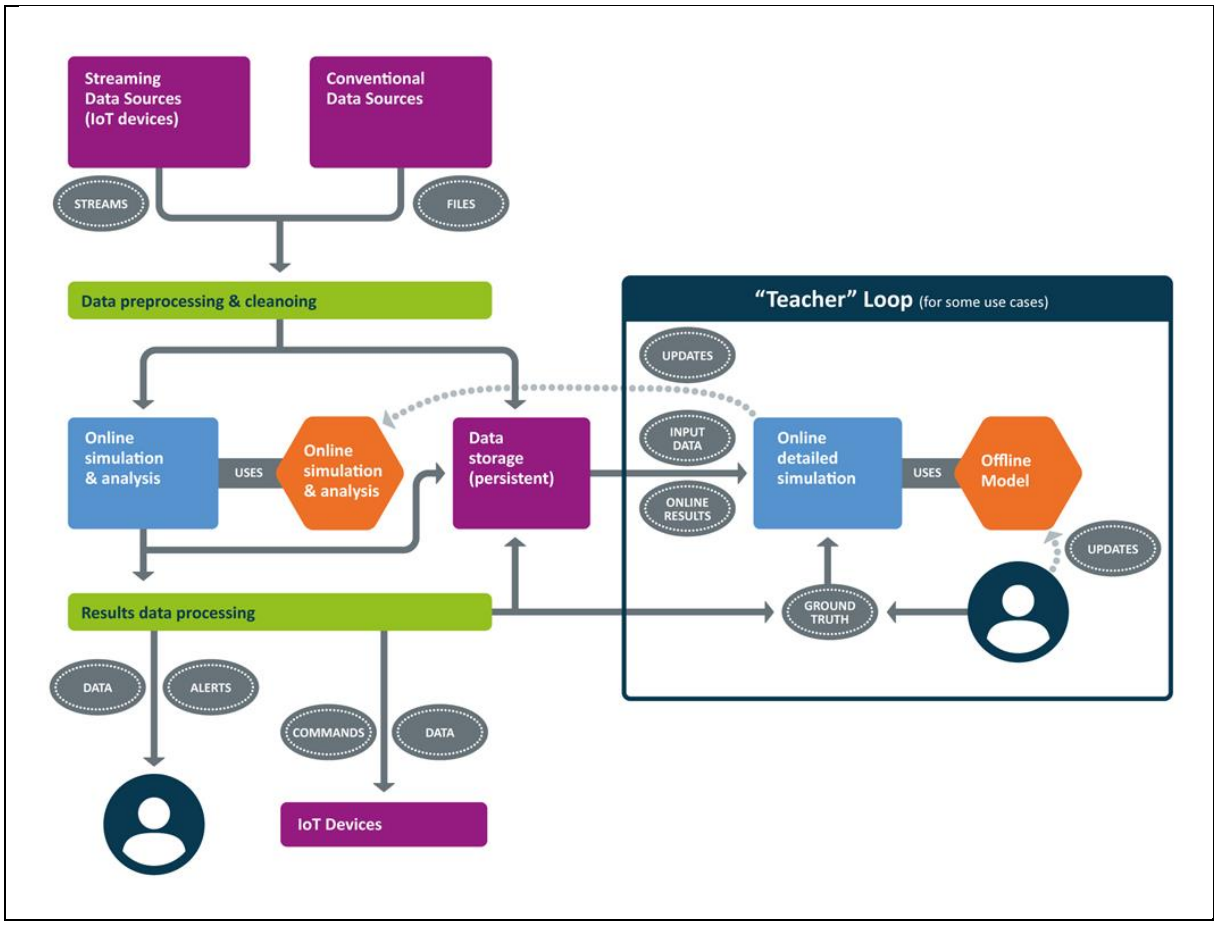

Figure 3: A typical mixed simulation and machine learning workflow

Figure 3 unfolds the loop shown in Figure 1 and shows three steps that are common to the use cases discussed: in a first step, data from a multitude of real-world sensors or conventional sources (e.g. databases) is ingested, pre-processed and cleaned. This already can involve significant processing, like in situations where the analysis of correlation between independent data streams is required. All or part of the resulting data is put into storage for documentation and for use in improving the analysis/simulation models.

The second step consists of an in-depth analysis of the data from step 1 – this includes anything from image classification to computing the next status of a complex technical twin using multi-discipline simulation techniques. The online model encodes the analysis steps, and it can range from a simple rule set to a complex HPC simulation code. Part of the analysis results are again put into storage for later use.

The third step is the processing of the analysis results, and the communication with human users or IoT devices/Cyber-Physical-Systems (CPS). Depending on the nature of the problem, the loop can be closed by the commands passed to a CPS effecting its sensor readings, which requires the update of the analysis in step 2. In the Digital twin case, the analysis in step 2 keeps its own state and runs in 'streaming mode', receiving updates from the real world, reconciling them with the CPS's virtual model, and sending out commands to the CPS.

The role of the 'teacher loop' is very apparent for Deep Learning based analysis approaches – the online model at the heart of step 2 is created in a separate training phase and then made 'live'. For reinforcement learning, the online model is improved by assessing its performance and rewarding/punishing certain aspects. Taken to the next step, the online model could represent a simplified version of a car (for example), which is updated and extended/improved

by a full, physically correct car model. The key idea behind splitting off the teacher loop is that it can be disconnected after a while (analogously to real life with teachers and pupils, once certain proficiency has been achieved). The online model can then be made significantly simpler than, for instance, a fully physically correct six degrees of freedom driving model, reducing the amount of processing needed per instance and consequently reducing energy requirements.

1.3 Examples of use cases spanning across the digital continuum

We introduce two sources of uses cases:

- The 'Extremes prediction' use case in the context of the 'Destination Earth' initiative of the EC (ECWMF)
- Use cases on DTs presented by EU-MATHS-IN (areas: industrial, health, environmental)

1.3.1 *Extremes prediction*

Introducing the topic

Digital Twins (DT) originate from industrial production control processes whereby a digital replica of the actual process receives real-time information from observations of the physical system to optimise production, adapt to external factors and enhance resilience by predicting component failure.

This concept is highly relevant for industries closely tied to digital technologies, but it can be equally applied to use cases outside industrial production. Even though high-performance computing (HPC) and big-data (BD) handling are at the core of DTs, only the full integration of all components of the digital continuum will allow to exploit the full potential of both HPC and BD and generate effective DT capabilities for all use cases.

Example use case Extremes Prediction

Natural hazards represent some of the most important socio-economic challenges our society is facing in the next decades. Natural hazards have caused over 1 million fatalities and over 3 trillion Euros economic loss worldwide in the last 20 years, and this trend is increasing given drastically rising resource demands and population growth. Apart from the impact of natural hazards on Europe itself, the increasing stress on global resources will enhance the political pressure on Europe through yet unprecedented levels of migration.

Dealing responsibly with extreme events does not only require a drastic change in the ways our society is solving its energy and population crises. It also requires a new capability of using present and future information on the Earth-system to reliably predict the occurrence and impact of such events. A breakthrough of Europe's prediction capability can be made manifest through science-technology solutions delivering yet unprecedented levels of predictive accuracy with real value for a society.

These science-technology solution includes the entire loop from (1) basic Earth-system science, established in (2) enhanced prediction models significantly, combined with (3) the vast range of Earth observation data ranging from advanced satellite instruments to commodity devices available through the internet of things, exploiting (4) extreme-scale computing, big-data handling, high-performance data analytics and artificial intelligence technologies, for feeding (5) impact models that translate scientific data into information close to the real users responsible for critical infrastructures, hydrology and water, energy, food and agriculture, health and disaster management.

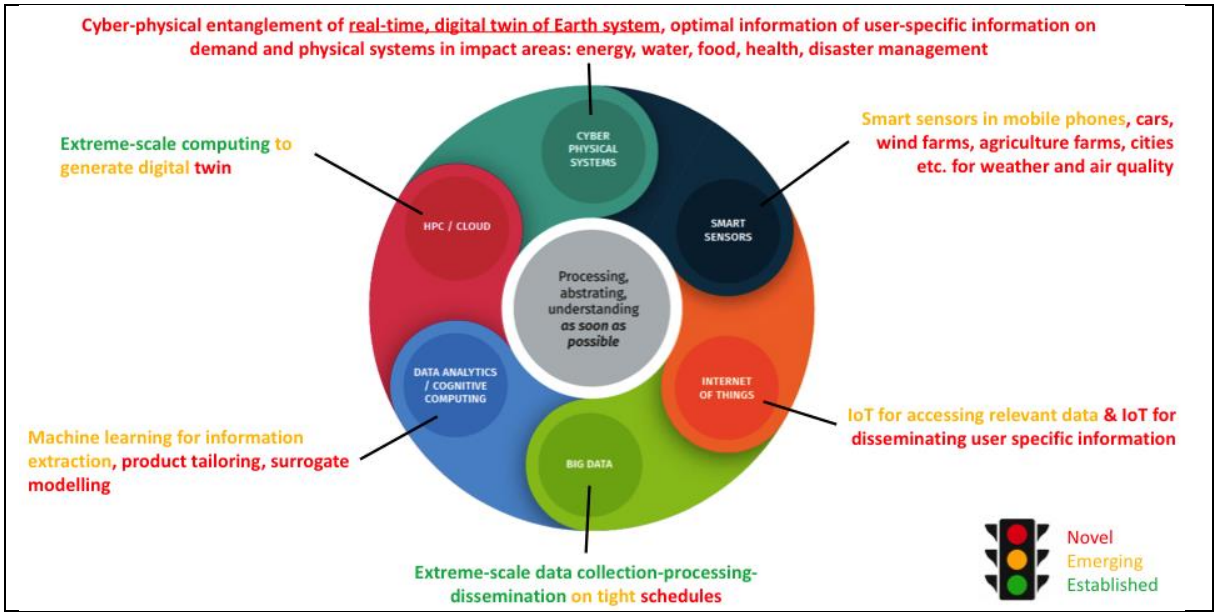

Figure 4: Main elements of digital continuum and relevance for extremes prediction use case

Figure 4 shows the main elements of the digital continuum and the level of sophistication presently available to this use case. Note that all elements need significant research and innovation support.

Research challenges:

Smart sensors and Internet of Things: At present, Earth-system observations already comprise hundreds of millions of observations collected daily to monitor atmosphere, oceans, cryosphere, biosphere and the solid Earth, the largest data volumes being provided by hundreds of satellite instruments. This volume is expected to increase by several orders of magnitude in the next decade, with a need to ingest such observations in digital twin systems within hours. Smart-sensor technology is highly relevant for satellite-based observations and dedicated station networks, but also for observations from commodity devices deployed on e.g. phones, car sensors and specialized industrial devices monitoring agriculture, renewable energy sources and infrastructures – made available through the internet of things. Such technology will allow outsourcing data pre-processing to Edge and Fog computing, and thus implement fully agile data management and information extraction.

Research and innovation is mostly needed for (1) smart-sensor design from IoT devices to satellite instruments that provide adaptive (targeted), continuous and quality controlled measurements of the environment plus all relevant meta-data to trace data origin and quality throughout the production process; (2) data compression and smart-selection procedures to maximize information content and access speed; and (3) open and real-time access to data.

Big data and data analytics: The daily volume of Earth-system observation and simulation data already exceeds petabytes today, prohibiting effective and timely information extraction, critical for proactive and reactive response for anticipating and mitigating the effects of extremes. Both simulations and observations need to be generated and assimilated in the Earth-system's digital twin within minutes to hours of time-critical workflows towards near-real-time decision-making. Overcoming the data-transfer bottlenecks between the digital twin and downstream applications is crucial and future workflow management needs to make such applications an integral part of the observation and prediction infrastructure. Powerful data

analytics technology and methodologies offer the only option to make the effective transfer between raw data and information tailored to those sectors needing to prepare and respond to extremes, namely water, food, energy, health, finance and civil protection.

Research and innovation is mostly needed for (1) data-centric and resilient workflows with powerful object-storage exploiting deep-memory hierarchies; (2) pooling of applications which have data dependencies; (3) machine learning (ML)-based input/output data selection, quality control and error correction methods; and (4) ML-based data compression and feature detection methodologies.

High-performance and Cloud computing: Today, experimental and operational Earth-system simulations use petascale HPC infrastructures, and the expectation is that future systems will require about 1000 times more computational power for producing reliable predictions of Earth-system extremes with lead times that are sufficient for society and industry to respond. This need translates into a new software paradigm to gain full and sustainable access to low-energy processing capabilities, dense memory hierarchies as well as post-processing and data dissemination pipelines that are optimally configured across centralized and Cloud-based facilities. European leadership in this software domain offers a unique opportunity to turn the European investment in HPC digital technology into real value.

Research and innovation is mostly needed for (1) application co-design: interactive workflow management combining physical models, observational data handling and ML methods; (2) mathematical methods and algorithms: combination of higher order, grid-point discretisations with large time stepping methods, parallel-in-time algorithms, multi-grid/level solvers, ML-based surrogate models, observational data assimilation and methods for data analytics, mixed arithmetic precision (including non-IEEE); (3) I/O and storage: deep-memory infrastructures, efficient file systems, I/O servers and objected based data stores; and (4) programming environment: high-productivity language environments and DSL tool-chains to create back-ends optimized for heterogeneous processors, performance models and optimization tools. But hardware system development is also highly relevant for this use case, namely heterogeneous processor configurations in fat-nodes (including data-flow engines), high-bandwidth upper-memory layers and deeper memory hierarchies for data processing on the fly, and configurable computing.

Centre-to-edge framework: At present, the main computing tasks are performed on centralized, dedicated systems where also the main data storage facilities lie. The observational input data flow crosses all levels of edge, cloud and centralized computing during which selected pre-processing steps are exercised, for example for satellite collection and pre-processing at distributed receiving stations of the ground segment, their further dissemination and processing by space agencies and meteorological centres, and assimilation into models at prediction centres. Similar data flow mechanisms exist for ground-based meteorological observation taken from networks, stations and (few) commercial providers. For IoT devices, this infrastructure does not yet exist. Increasingly, the back-end of model output data processing interfacing with commercial service providers is being placed into the cloud also offering better access to ML-based data analytics.

Research and innovation is mostly needed for (1) interoperable ML methodology applied to the entire data flow for observational data for better quality control and data selection; (2) dedicated, smaller sized clouds for sharing HPC workloads without losing processing speed; (3) fast access to streaming devices with standardized processing tools; and (4) open access mechanisms for commercial data from IoT devices.

Inter-connected digital infrastructure: Collecting and transferring massive amount of data among end devices scattered in multiple areas, aggregation points, centralized digital platforms and computing centres, requires a suitable underlying network infrastructure. Evolved networks and services target to offer secure and trustable solutions that will support the desired QoS for different flows. These networks will maximize the usage of European HPC capacities (e.g., connect edge pre-processing centres to the main core DT centres). Cross-European data-intensive applications, exploiting the foreseen computing infrastructure, will entail the existence of fast, secure and reliable networks. In this sense, the interconnecting networks have the potential to become an integral part of the critical and strategic HPC infrastructure, provided that synergic efforts and investments are realised. Smart Networks and Services emerge as a key enabler for the transformation in the HPC domain.

At the same time, HPC is expected to play an important role in the evolution of communication networks. As several use cases require close to zero latency, the extensive use of edge computing need to rely on low cost yet high-performance computing facilities that will interact with end devices and also among themselves. Similar needs have been identified for use cases where closed local networks are needed by specific vertical industries. Finally, HPC solutions are envisioned as an integral part of future networks to orchestrate and dynamically manage computing resources needed by different network slices. Thus, Smart Networks and Services are expected to use HPC solutions to improve their performance.

Research and innovation are needed in the fields of (1) advanced traffic steering solutions over multiple radio links (including terrestrial and satellite) that will offer secure and reliable end-to-end data paths through tighter inter-working among services and network components; (2) Seamless fog/edge/cloud computing orchestration; and (3) Self-reacting core orchestrators, with capacity for event detection, reconfiguration decision, deployment and operation.

1.3.2 *Real time simulation, lifetime monitoring, and control of complex models*

Introducing the topic

Digitalisation changes everything everywhere. With the rise of new technology trends, such as AI Foundations, Intelligent Things, Cloud to Edge, or Immersive Experiences, many of today's paradigms can be expected to be disrupted. A Digital Twin consistently and in real time integrates all data (test and operation data), models (design drawings, engineering models, analyses), and other information (requirements, orders, inspections) of a physical asset generated along its life cycle into a computer simulation model to predict and optimize performance and maintenance. Digital Twins are the next wave in simulation technologies, merging high-performance simulation with big data and artificial intelligence technologies. It is predicted that companies who invest in digital twin technology will see a 30% improvement in cycle times of critical processes [4].

Figure 5: Simulation is evolving from a trouble shooting tool to key business

Many components of the physical model-based digital twin vision are not new. For almost all domains of science and engineering and in almost all industrial sectors, model-based approaches are well-established. A multitude of commercial and open-source software for modelling, simulation and optimisation in data-rich environments (MSODE, [13]) based on mathematical methods is available. All this is fostered by computers becoming more and more powerful.

However, the advancements in the different fields alone are not enough to achieve construction and application of real time simulations with digital twins. Currently, due to the high manual human effort for integration of the various approaches, only major companies with large R&D departments will afford to build digital twins, but it would be desirable that companies on all scales profit from the development. In particular, novel mathematical and computer technologies are required to describe, structure, integrate, and interpret across many engineering disciplines.

Technology Challenges for real time simulation, lifetime monitoring, and control of complex models

Currently, models, methods, as well as software implementations and data sets are of highly different fidelity requiring many manual interactions. To meet the future challenges, it is necessary to develop novel MSODE paradigms with a systematic MSODE-based approach that allows to build highly automated modularized networks of model hierarchies (from very high-fidelity physics based models to very coarse, surrogate, or even purely data based models),

which can deal with multi-physics and multi-scale systems. Key to this will be a convergence of Artificial Intelligence methods and first principle approaches typically used in MSODE by laying down novel mathematical principles as the core language of digital twins. The role of advanced and multi-fidelity modelling is equally fundamental in the development of current Edge Computing ambitions, in order to deliver new services providing the right information at the right place in an efficient way.

Furthermore, the model hierarchies should:

- be able to (automatically) evolve with the availability of new information, data, or even changes in the process,
- allow adaptive models and solutions with seamless choice of accuracy and speed, in particular allow real-time and interactive simulation and optimisation,
- be able to quantify the uncertainties and risks that come with the determined solutions and in particular be made robust towards inaccuracies in the data and the model,
- exploit new computing architectures, e.g. combined cloud - edge solutions,
- be flexible for new user interaction concepts, to enable also non-experts to benefit from digital twins.

Possible use cases include, among others, digital twins for industrial assets, production systems, energy networks, electrical circuit design, urban air pollution, and personalized therapies.

There are natural links of the theme of real-time digital twin applications to existing Centres of Excellence running by the projects ChEESE [13], EoCoE-II [14], ESiWACE2[15], EXCELLERAT [16], HiDALGO [17] and mission areas like climate-neutral and smart cities.

Research challenges

Modelling, Simulation and Optimization in Data-rich Environment (MSODE): Implementation of online simulations of physical assets and processes is possible only with the model order reduction (MOR). Based on mathematical and engineering models, computational complexity can be significantly diminished via combining simulations and machine learning (ML), which enables even 10.000 times faster simulation on industrially relevant assets on the price of slight accuracy losses. The fast simulations will be exploited by tailored methods for control, optimisation and uncertainty quantification. HPC and Big Data technologies are key elements in MOR, namely in the context of information gathering on the states and information extraction. For more details, see the MSODE SRA [18].

Research and innovation is mostly needed for (1) further development of model order reduction methodology and implementations, in particular for large scale computational mechanics and fluid dynamics models; (2) stochastic and robust distributional optimization, robust control theory; (3) model and ML-based techniques for missing data; (4) grey box models combining data-based and machine-based learning techniques; (5) real-time model predictive control and reinforcement learning; (6) explainable reinforcement learning and model predictive controls; (7) large scale model predictive control with mixed (continuous and integer) variables under PDE constraints; (8) optimisation algorithms to minimize the objects‘ misclassification in ML; (9) port-Hamiltonian hierarchical modelling of systems; (10) high automation for control generation with minimum expert interaction; and (xi) data assimilation for complex high dimensional models.

Smart sensors and Internet of Things: The concept of many novel engineering applications relies heavily on smart sensors and the IoT. However, in many cases the number of sensors is

limited or simply not available. hindering the realisation of smart IoT applications. Real-time predictive digital twins allow realising virtual sensing capabilities, where virtual sensor values are inferred from the digital twin in combination with real sensors.

Research and innovation is needed to (1) provide accelerated simulation techniques, e.g. through concepts of model order reduction; (2) appropriate estimation techniques for their corresponding parameters, as well as advanced uncertainty quantification technologies to determine the reliability and accuracy of digital twins; and (3) domain specific IoT sensing (e.g. in healthcare industry for human organ sensing) interfaced with digital twins.

Big data and data analytics: Data analytics has been shown a successful concept in many consumer as well as industrial applications. On the one hand, data is available abundantly in many consumer applications, it is rather limited in many industrial applications. While on the other hand, we have no mechanistic understanding of human behaviour in many situations, our knowledge of natural science as well as industrial assets and systems is rather deep. However, ways to reuse corresponding physical models in data analytics approaches are not available.

Research and innovation is mostly needed for (1) the combination of data analytics methods and mechanistic mathematical models, e.g. introducing partial differential equations as regularisers as introduced in the concept of Physics Informed Neural Networks (PINNs); (2) systematic mathematical and algorithmic integration of machine learning and classical modelling, simulation and optimization technologies, which will be a tipping point for many industrial applications; (3) efficient implementation of dynamic neural networks with rapidly varying data infusion; (4) purely machine learning based methods for large-scale black-box simulations; and (5) robustness analysis of ML algorithms.

High-performance and Cloud computing: With the growing understanding of our technical systems, the required computational power to simulate the systems is exploding. Standard workstations are not enough anymore to compute these systems in acceptable time frames. However, just migrating to HPC systems might not be enough.

Research and innovation is mostly needed for (1) mathematical and algorithmic research to split up system models into decomposable models which can be split in a pre-computable offline phase and an efficiently computable online phase. The pre-computable phase will still require high-performance and cloud computing infrastructures to be computed; (2) Scalable HPC simulation of hierarchic models with complex dataflow; (3) automatic generation of online digital twins via high -performance computing (HPC); and (4) fast solution of extremely large scale sparse linear systems.

Centre-to-edge framework: With control and operational 'intelligence' moving towards edge applications, predictive simulations will need to be performed on the edge combined with data analytics applications.

Research is required for (1) efficient split up computations between computational heavy offline phases and computationally online phases (c.f. above) with novel mathematical concepts and algorithms; (2) evaluation of the differentiation between offline pre-computations and online computations that might be smeared with the advancement of 5G; (3) investigation of strong computational performance on premise or in the cloud with low latency might be available; (4) efficient mathematical algorithms in such heterogeneous systems, to rethink many of today's paradigms; and (5) real-time predictive twins for edge devices and PLC.

2 Collaboration Initiatives focussing on the digital Continuum

2.1 The starting point: comparing compute stacks

In 2015, the HPC Community represented by ETP4PC and the Big Data Community represented by the Big Data Value Association held a joint work session which provided an opportunity for an interdisciplinary discussion regarding the current strengths of and differences between the (software and hardware) stacks of Big Data Computing and HPC. The main question raised was how the current strengths of one stack may address a shortcoming or need in the other stack and vice-versa. A white paper [5] was jointly issued to document this initiative.

In 2017, as a result of a further joint analysis of use cases, this position was extended to include Deep Learning as shown in Figure 6. To a large extent, this work was also influenced by the results and insights obtained in the work sessions run by the BDEC initiative [1].

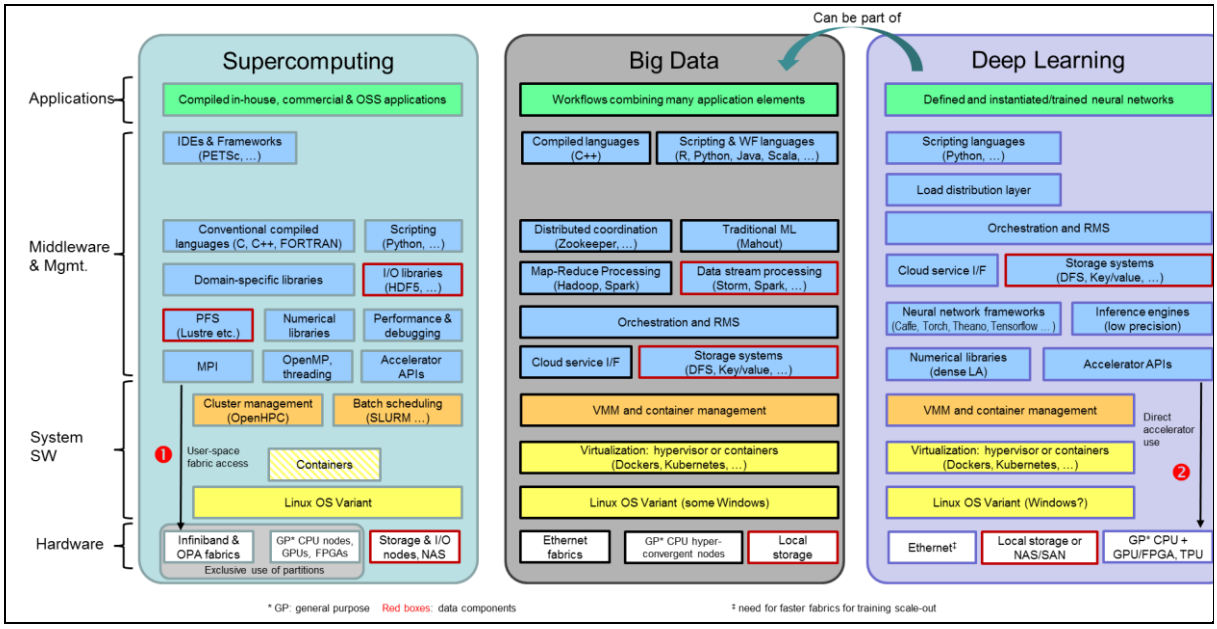

Figure 6: The three compute stacks side-by-side

In the first quarter of 2020, the 4th issue of the ETP4HPC Strategic Research Agenda (SRA-4) was released under Task 2.1 of EXDCI-2 [4]. The notion of 'HPC in the loop' (see Figure 1) was introduced to connote the interaction of several disciplines which eventually triggered the formation of the TransContinuum Initiative (TCI), as further outlined in the following chapter.

2.2 The TransContinuum Initiative (TCI)

By July 2020, seven associations had agreed to establish the 'TransContinuum Initiative' (TCI):

- ETP4HPC - European technology Platform for High Performance Computing [6]
- BDVA - Big Data Value Association [7]
- ECSO - European Cyber Security Organisation [8]
- 5G IA – 5G Infrastructure Association [9]

- Claire - Confederation of Laboratories for Artificial Intelligence Research in Europe [10]
- EU MATHS IN - European Service Network of Mathematics for Industry and Innovation [11]
- HiPEAC - High Performance Embedded Architecture and Compilation [12]

A jointly developed vision document first introduces the objectives of this horizontal collaboration:

Figure 7: The charter of the TransContinuum Initiative (TCI)

The Digital Continuum

We observe a continuous miniaturisation of computing and storage devices as well as their ubiquitous deployment on the one hand ranging from data centres to the edge and beyond. On the other hand, we also see the need for enabling workflows beyond single control domains like a data centre. Therefore, a new overall system architecture design to accommodate the ecosystem change is to be expected in the coming decades (environmental and technological) and to integrate the different actors horizontally. The new demands and challenges that combine data, storage and compute, distributed across the continuum, and the maintenance and resource efficiencies, are pushing for drastically increased software and hardware *sustainability*. Furthermore, the need to provide high-level *cybersecurity* is profoundly changing the game. Efficiency and resilience will have to reach levels never achieved so far, while taking into account the intrinsic distributed and heterogeneous nature of the continuum. In addition, the question of dealing with very high volumes of data needs to be faced, and the preponderance of quality versus quantity will become unavoidable. These considerations will spread over all components. Long-lifetime hardware devices will have to be reconfigurable, modular, exchangeable, and self-aware in order to be operational over extended periods. Algorithmic efficiency will need to be drastically improved (e.g. more efficient AI), which requires development of basic modelling, simulation and optimization methodologies in data-rich environments (MSODE), including model-order reduction. Management and deployment of large-scale application workflows will have to be adapted or invented. Network protocols will have to offer better control over the data logistics.

Furthermore, it is widely recognised that AI will play a central role in these extreme-scale, continuum infrastructures. This will occur at three levels:

- AI for Digital Infrastructure,
- Digital Infrastructure for AI, and
- AI for Science, Industry and Societal Challenges.

The first level addresses how AI inspired techniques can pilot and monitor the continuum and in doing so provide solutions to the points listed in the previous paragraph. The second level treats the question of re-designing the e-infrastructure to efficiently deal with data analysis and machine learning, which means, among others, tuning of data access, I/O, low precision arithmetic, and moving code and data where it will be the most efficiently performed. The last level deals with the ever-increasing needs to exploit AI techniques for extreme-scale, combining data and compute through the interpretation and coupling of computing results, measurements and observations (e.g. Digital Twins in extreme earth modelling, combining climate models with satellite data and on-ground sensors).

The overall objective is to target high TRL solutions (7 or more), based on horizontal synergies between all the concerned digital infrastructure technologies: HPC, Big Data, Machine Learning, IoT, 5G, cybersecurity, processor technology (EPI) and robotics. All of these components of the digital infrastructure *together* will be able to address the critical societal challenges and sustainable development goals by mobilising their amazing potential all the way across the continuum.

The TC-Initiative will focus on collaboration in the following five areas:

- Identify priorities and recommendation for European R&I work programs - Jointly we will elaborate recommendations for R&D to be carried out in EU-funded work programs addressing challenges in the digital continuum. The recommendations will cover challenges in technological (hardware and software) functionality, interoperability, and APIs. New standards, best practices, methodologies and project-type related suggestions will also be generated. Applications deployed in the digital continuum are addressed wherever needed.
- Interlock with European R&I funding agencies and R&D programs (e.g. JUs, Missions) - The recommendations will be presented to EU-funding entities like Joint Undertakings (JUs) and applicable programs in the MFF 2021-2027. TCI representatives will be available for presenting and explaining the recommendations as well to discuss any possible further analysis and elaborations.
- Generate and foster an interdisciplinary network of experts - We look forward to a lively exchange of news about EC work programs, calls and related events, events of partner organisations and potentially joint activities. We will jointly analyse new industrial and scientific use cases to better understand the challenges presented. On one side, this is a pre-requisite for any R&I recommendations, on the other side it facilitates the forming of interdisciplinary consortia for upcoming calls.
- Contribute to SR(I)As and other partners' documents - Based on the results of the joint work mentioned above, contributions to the Research and Innovation Agendas or any other road mapping documents issued by participating partners will be offered.
- Contribute to the five Horizon Europe missions - One of the first pragmatic actions will be to design the contribution of the Digital Twin enabler to the Horizon Europe missions (adaptation to climate change including societal transformation, cancer, healthy oceans, seas coastal and inland waters, climate-neutral and smart cities, soil health and food.) These missions will need digital technologies to achieve their respective goals and Digital Twins should be one of the key elements.

3 Conclusion and outlook

The deployment of 'High-Performance Computing' is undergoing a significant change. 'HPC' does not apply to only supercomputers in large datacentres but also to other scales, such as embedded or edge deployment. Furthermore, HPC now finds itself at the heart of a compute infrastructure supporting simulation, modelling and data analysis in a digital computing continuum. Core HPC technologies and methodologies are being used to enable concurrent processing to permeate all levels of that digital computing continuum.

Research on both HPC applications as well as on HPC technology will expand from the current fields deploying HPC solutions to adjacent fields to address AI, Data Analytics and IoT-related challenges. This will influence the selection and definition of research priorities, which can only be effective as the result of a true interdisciplinary effort.

The TransContinuum Initiative is a promising start in the direction of facilitating a collaboration between European associations and projects interested in strengthening the digital infrastructure that is essential for the implementation of many European priorities such as the 'Green Deal', the Horizon Europe Missions and the Destination Earth project.

The next task ahead, far beyond the lifetime of EXDCI-2 project, is to analyse industrial and scientific use scenarios reflecting the Digital Twin concept. The outcome of this work should be used to identify the building blocks of the next recommendations for the European R&D priorities.

4 Annex: Actions in the context of the TransContinuum Initiative

April 2018	Conducted an 'extreme compute – extreme data' use case analysis workshop with technical experts from BDVA, Exdci-2 and external HPC experts. Out of 15 use cases, 5 were analysed further to understand the diversity of the two technology stacks. A joint BDVA - ETP4HPC document was generated discussing the options for convergence of the two stacks and their basic characteristics and usage patterns. See: https://www.etp4hpc.eu/bigdata.html
June 24th, 2018	Work session with international contribution. Focus was a vision for the next 5 years for HPC technology and use evolution. The 5 presentations led to a second workshop on June 27th with the SRA technical experts and working group leaders to identify the main drivers for the next generation of research priorities and the implications with adjacent domains like AI, Data Analytics and IOT.
December 2018	At ICT 2018, BDVA and ETP4HPC held a session on 'HPC & Bid Data' on the digital continuum between HPC centres, cloud, flog and edge computing.
December 2018	At the EBDV-Forum : participation in a panel discussion on the collaboration necessity between European associations promoting HPC, Big Data, Cyber Security, Robotics, 5G and IOT. Triggered by this event a closer collaboration between ETP4HPC, ECSO and AIOTI started and the joint work with BDVA was intensified.
December 2018	A MoU with BDVA and one with AIOTI was signed focussing on joint road mapping work for future SRAs.
May 17th, 2019	Conducted an EXDCI-2 SRA work session with technical experts, as part of the EuroHPC HPC Summit week. New working group leaders came on board and a consensus was reached on how to qualify research priorities for the SRA. Also, several industrial use cases were presented and examined following the approach outline in the Blueprint paper
May 24th, 2019	Started the set of 8 working group kick-off calls (in total with over 100 working group members from industries, academia, research centres and organisations like HiPEAC, BDVA and AIOTI.
December 2019	The SRA-4 document was finalised. (https://www.etp4hpc.eu/pujades/files/ETP4HPC_SRA4_2020_web(1).pdf)
January 2020	On the basis of the concept for R&I goals in support of the Transcontinuum laid out in the SRA-4, the process of generating interest to form a horizontal partnership was started with BDVA and HiPEAC
March 2020	HiPEAC and BDVA agreed to collaborate, ECSO joined as well
April 2020	With EU-MATHS-IN and 5G-IA two additional associations joined. A series of telcos was held to discuss the vision, scope, goal and set-up of the collaboration
May 2020	Several dedicated interdisciplinary working groups were formed, the interest in the associations for this initiative increased significantly.
June 2020	Several tutorials were given during webinars and dedicated small group focus sessions.
July 2020	The 2-page vision document (see chapter 2.2 above) was agreed upon and finalised.

Table 1: Working sessions and workshops evolving the TransContinuum Initiative