

SPECTRUM

D6.1 Technical Blueprint for Compute and Data Continuum

Status: Approved by the EC

Dissemination Level: Public



Funded by
the European Union

Disclaimer: Views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them. SPECTRUM is funded by the European Union - Grant Agreement Number 101131550 - www.spectrumproject.eu

Abstract	
Key Words	High Performance Computing, HPC, High Energy Physics, HEP, Radio Astronomy, RA, compute and data continuum, storage, network, European Union, EU, supercomputing, AI, ML, quantum computing
This document presents a Technical Blueprint for a compute-and-data trans-continuum infrastructure design and architecture optimised for cost, power efficiency, time to results, and science-driven capability.	

Document Description			
D6.1 Technical Blueprint for Compute and Data Continuum			
Work Package Number 6			
Document Type	Report		
Document status	Approved by the EC	**Version**	2.0
Dissemination Level	Public		
Copyright status	 This material by Contributing Parties of the SPECTRUM Consortium is licensed under Creative Commons Attribution 4.0 International License.		
Lead Partner	CERN		
Document link	https://documents.egi.eu/document/4074		
Digital Object Identifier	https://zenodo.org/records/20138626		
Author(s)	Lead Editors: • Eric Wulff (CERN) • Jeff Wagg (CNRS) • Maria Girone (CERN) • Amael Ellien (CNRS) Contributors: • Andrea Manzi (EGI Foundation) • Sergio Andreozzi (EGI Foundation) • Hans-Christian Hoppe (FZJ) • Luis Cifuentes (FZJ) • Shan Mignot (CNRS) • David Southwick (CERN)		
Reviewers	Tommaso Boccali (INFN) John Swinbank (NWO-I ASTRON) Raymond Oonk (SURF)		
Moderated by	Patricia Ruiz (EGI Foundation)		
Approved by	AMB		

Revision History			
Version	**Date**	**Description**	**Contributors**
V0.1	06.11.2024	Created draft of document outline.	Eric Wulff
V0.2	18.03.2025	First description of the key technical activities.	Eric, Maria Girone
v0.3	15.04.2025	Added section covering preliminary work done in T5.1.	Eric Wulff, David Southwick
v0.4	20.04.2025	Defined the first list of building blocks.	Eric Wulff
V0.5	22.07.2025	First detailed description of all building blocks.	Eric Wulff, Andrea Manzi
V0.6	02.09.2025	Added section covering preliminary work done in T5.2 and T5.3.	Hans-Christian Hoppe, Luis Cifuentes
V0.7	13.10.2025	Changed building blocks to capabilities and created a capability map. Added more perspectives from RA.	Eric Wulff, Jeff Wagg
V0.8	11.11.2025	Added section on sustainability	Shan Mignot
V0.9	09.12.2025	Completed internal review.	Tommaso Boccali, John Swinbank, Raymond Oonk, Luis Cifuentes, Hans-Christian Hoppe, Maria Girone, David Southwick, Jeff Wagg, Eric Wulff
V1.0	12.12.2025	First public draft	Eric Wulff, Jeff Wagg, Maria Girone, Andrea Manzi, Shan Mignot, Tommaso Boccali, John Swinbank, Raymond Oonk, Luis Cifuentes, Hans-Christian Hoppe.
V1.1	29.04.2026	Revised version following input from external experts and internal reviewers	Eric Wulff, Jeff Wagg, Maria Girone, Andrea Manzi, Shan Mignot, Tommaso Boccali, John Swinbank, Raymond Oonk, Luis Cifuentes, Hans-Christian Hoppe, Amael Ellien.
V1.2	07.05.2026	AMB Approval	
V2.0	15.05.2026	**Final**	

Terminology / Acronyms	
Terminology / Acronym	**Definition**
AI	Artificial Intelligence
AAI	Authentication and Authorisation Infrastructure
ATLAS	A Toroidal LCH Apparatus
AARC	Authentication and Authorization for Research and Collaboration
BPA	Blueprint Architecture
CERN-SSO	CERN Single Sign-On
CMS	Compact Muon Solonoid
CoP	Community of Practice
CPU	Central Processing Unit
CUDA	Compute Unified Device Architecture
CVMFS	CERN Virtual Machine System
DAG	Directed Acyclic Graph
DOI	Digital Object Identifier
EFP	EuroHPC Federation Platform
EOSC	European Open Science Cloud
ESCAPE	European Science Cluster of Astronomy and Particle Physics ESFRI research infrastructures
FAIR	Findable, Accessible, Interoperable, Reusable
FLOPS	FLoating-point Operations Per Second
FPGA	Field Programmable Gate Array
FTS	File Transfer Service
GDPR	General Data Protection Regulation
GHG	GreenHouse Gas
GPU	Graphics Processing Unit
HEP	High Energy Physics
HL-LHC	High-Luminosity LHC
HTC	High-Throughput Computing
HPC	High-Performance Computing

HPCG	High Performance Conjugate Gradients
IAM	Identify and Access Management
Indigo-IAM	INDIGO DataCloud Identify and Access Management
I/O	Input/Output
IT	Information Technology
JU	Joint Undertaking
KPI	Key Performance Indicator
LCA	Life-Cycle Assessment
LHC	Large Hadron Collider
LOFAR	LOw Frequency ARray
MFA	Multi-Factor Authentication
ML	Machine Learning
QCD	Quantum Chromodynamics
RA	Radio Astronomy
RI	Research Infrastructure
SKA	Square Kilometre Array
SKAO	SKA Observatory
SRCNet	SKA Regional Centre Network
SRIDA	Strategic Research, Innovation and Deployment Agenda
SSH	Secure Shell
SSL	Secure Sockets Layer
SSO	Single Sign On
VO	Virtual Observatory
VOMS	Virtual Organisation Membership Service
VM	Virtual Machine
WG	Working Group
WLCG	Worldwide LHC Computing Grid
WP	Work Package

Table of Contents

List of Figures **8**
List of Tables **8**
Executive summary **9**
1. Introduction **11**
1.1. Background 11
1.2. Methodology 13
1.3. Related initiatives 14
2. Preliminary work **15**
2.1. The SPECTRUM Community of Practice 15
2.1.1. The Knowledge Hub 15
2.1.2. Findings 15
2.2. Representative use cases 16
2.2.1. Key Findings and Recommendations 17
2.3. Landscape, Requirements and Gap Analysis 18
2.3.1. European e-infrastructure landscape 18
2.3.2. Proposed requirements 19
2.3.2.1. Storage/Data Capabilities and Services 19
2.3.2.2. Data Transfer and Federation 20
2.3.2.3. Compute Capabilities and Services 20
2.3.2.4. Software Services, Interfaces, and Stacks 21
2.3.2.5. Compute Federation Services 21
2.4. Interoperable access policies 22
2.4.1. Access Policies, Gaps and Recommendations 22
2.5. Consolidated Gaps/Requirements and Recommendations 24
3. Sustainability **28**
3.1. Data and software preservation and reuse 28
3.2. Environmental sustainability 28
4. The Compute and Data Continuum **34**
4.1. Compute Resources 34
4.2. Data Resources 36
4.3. Software Distribution and Execution 37
4.4. Orchestration and Workflows 38
4.5. AI/ML and HPC Applications 39
4.6. Resource Federation 40
4.7. Monitoring and Observability 40
4.8 Security and Trust 41
5. Technical activities and recommended actions **43**
5.1. Standardization of Interfaces 43
5.2. Co-Design of Computing, Data, and Networking Infrastructure 44
5.3. Software Portability and Heterogeneous Architectures 45
5.4. Data Management and Network Performance 46

5.5. Workflow Adaptation and Optimization 47
5.6. Security, Authorization, and Authentication 48
5.7. AI/ML Integration and Computational Trends 49
5.8. Long-Term Resource Provisioning and Accounting 49
6. Conclusions **51**
Appendices **53**
A Related Initiatives **53**

List of Figures

- Figure 1: Capability map for the HEP and RA communities.

List of Tables

- Table 1: Consolidated gaps/requirements and recommendations from the study of the landscape of RIs, use cases from HEP, RA and beyond, as well as interoperable access policies.
- Table 2: Environmental impact categories as defined in the Product Environmental Footprint method recommended by the European Commission.

Executive summary

European research communities are preparing for decades of data growth and increasing computational complexity. Infrastructures such as the High-Luminosity Large Hadron Collider (HL-LHC) in High-Energy Physics (HEP), and the SKA Observatory and Low Frequency Array (LOFAR) in Radio Astronomy (RA) will generate unprecedented data volumes, and rely on highly sophisticated workflows for simulation, analysis, and offline processing. Meeting these demands requires more than incremental improvements to existing e-infrastructures; rather, it calls for a coherent compute and data continuum that links together high-performance computing, cloud platforms, edge devices, and large-scale data services in a way that is reliable and easy to use for the scientific communities.

This document brings together the scientific, technical, and policy needs of the HEP and RA communities into a single, cohesive framework. Our approach is to specify the core capabilities required for an interoperable compute and data ecosystem and to provide clear and actionable recommendations for Europe's future e-infrastructure designs and research funding programs. The blueprint draws upon work conducted within the SPECTRUM project, including insights gathered from the Community of Practice (CoP), via public surveys, interviews with relevant experts, use-case analyses, and assessments of the infrastructure landscape. Building on these findings, a map of interdependent capabilities has been defined, formalising the structural aspects of a compute and data continuum: resource federation; monitoring and observability; security and trust; AI/ML and High-Performance Computing (HPC) applications; orchestration and workflows; software distribution and execution; and compute and data resources. This Technical Blueprint draws on the needs of the HEP and RA communities, but is designed to be more widely applicable to a large number of data intensive sciences, with similar computational needs.

The analysis underpinning this blueprint shows that while many of the capabilities already exist, significant gaps remain. Interfaces between infrastructures are often inconsistent, making interoperability a persistent challenge. Co-design between scientific communities and infrastructure providers is not systematic, which risks creating mismatches between what is built and what is needed. Portability of software across diverse architectures is still limited, particularly as accelerators (such as GPUs) become more common. Workflow management tools are not yet capable of seamless execution across multiple sites, and provenance capture[1] is inconsistently applied. Trust and security mechanisms remain fragmented, slowing down cross-border and cross-domain collaboration. Finally, long-term sustainability, covering management of environmental impacts over the full lifecycle of the infrastructure, operational funding, and workforce development, remains a key challenge for communities whose experiments will span decades.

This blueprint is intended to motivate a coordinated European effort to address these issues. Interfaces and standards must be harmonised to allow resources to interoperate smoothly. Co-design processes are to be strengthened so that major scientific projects can influence infrastructure roadmaps. Software portability and execution environments need investment to ensure that applications can exploit emerging hardware efficiently. Workflow orchestration must evolve into a cross-facility capability with built-in resilience and provenance. A federated trust and security framework is to be adopted, building on existing identity federations but extending them to new services and disciplines. AI/ML must be integrated as a core capability. Finally, policies regarding the use and the long-term management of the software and hardware resources are essential for sustainability. Together, these elements define a federated, sovereign, and future-proof infrastructure that will empower not only HEP and RA, but also many other data-intensive domains to fully exploit Europe's scientific potential.

The blueprint sets out how Europe can move from today's fragmented landscape toward an integrated, long-term continuum. By clearly defining which capabilities are needed, and identifying the most pressing gaps, it outlines a practical path for connecting compute, data, software, and security into a coherent infrastructure that is able to address the challenges of future data-intensive research. The approach is designed to be both technically realistic and aligned with the operational requirements of HEP, RA, and other scientific domains that depend on reliable, large-scale digital infrastructures, with strong focus on environmental sustainability and energy awareness.

[1] Provenance capture refers to the systematic recording of all information needed to understand, reproduce, and validate how a piece of data or a scientific result was produced.

By addressing these priorities, the blueprint also directly supports European policy goals of FAIR and open data, digital sovereignty, sustainability, and strategic autonomy. It provides a practical pathway to infrastructures that not only serve big data communities such as HEP and RA but also strengthen Europe's position in the global landscape of data-intensive science.

1. Introduction

1.1. Background

The European scientific computing landscape is undergoing a major transition as communities prepare for the exascale era. High-Energy Physics (HEP) and Radio Astronomy (RA) are among the most demanding fields in terms of data and computation, and so they illustrate the scale of the challenge ahead. Therefore, we have selected them as clear examples of future needs, with all the considerations being extendable to virtually all data-intensive scientific domains. The LOFAR2.0 radio interferometer will begin operating in 2026, and is expected to produce over a 100 Pb of science data by the end of the decade following approximately 1 billion CPU hours. The High-Luminosity Large Hadron Collider (HL-LHC), expected to start operations in 2030, will generate data volumes far beyond any previous scientific instrument. The two largest experiments at the LHC, namely ATLAS (A Toroidal LCH Apparatus) and CMS (Compact Muon Solonoid), are forecasting requirements for disk storage of approximately 3 EB and 1.5 EB respectively over the next decade; at the same time, both forecast tape archival storage requirements of approximately 4-6 EB by the mid 2030s[2,3]. The Square Kilometre Array Observatory (SKAO) will produce continuous data streams requiring advanced data processing at data centers in the host countries, and during full operations distribute up to 700 PB/year of scientific data for subsequent analysis to a network of regional centers known as the SKA Regional Centre Network (SRCNet).

High-performance computing (HPC[4]) is set to become a core component for HEP and RA. HEP experiments have been using HPC resources for more than 15 years, with HPC now making significant contributions to both production and analysis pipelines[5,6,7,8,9]. The Worldwide LHC Computing Grid (WLCG) already comprises approximately 1.4 million cores and over 1.5 EB of storage across 170 sites (home.cern), and HPC resources exploited by experiments range from 3 sites integrated by ALICE and LHCb, to more than 10 sites by ATLAS and CMS. The upcoming High-Luminosity LHC and SKAO will only increase these demands.

The European e-infrastructure ecosystem has grown organically over several decades and today includes national, international, and domain-specific initiatives. The EuroHPC Joint Undertaking (JU) provides access to leadership-class supercomputing resources through a network of petascale and exascale systems. The European Open Science Cloud (EOSC) serves as a federated environment for data, services, and software, enabling cross-disciplinary collaboration and FAIR data access across European research communities. The EGI Federation coordinates high-throughput computing resources across multiple member countries and domains. Domain-specific initiatives such as the WLCG and the emerging SRCNet address the particular needs of their scientific communities through specialized architectures and governance structures.

[2]CMS Phase-2 Computing Model: Update Document, CMS Offline Software and Computing (2022) https://cds.cern.ch/record/2815292

[3]ATLAS Software and Computing HL-LHC Roadmap, The ATLAS Collaboration (2022) https://cds.cern.ch/record/2802918

[4]Here, we define HPC as computers which are typically clusters of interconnected processors that are used to perform large-scale calculations.

[5]Strategy for HPC Integration in WLCG/HEP, Maria Girone et al. (2025) https://zenodo.org/records/15032799

[6]US ATLAS and US CMS HPC and Cloud Blueprint, Fernando Barreiro Megino et al. (2023) https://arxiv.org/abs/2304.07376

[7]Integrating the Perlmutter HPC system in the ALICE Grid, Sergiu Weisz et al. (2025) https://doi.org/10.1051/epjconf/202533701107

[8]Integrating LHCb workflows on HPC resources: status and strategies, Federico Stagni et al. (2020) https://doi.org/10.1051/epjconf/202024509002

[9] Bard, D., Benjamin, D., Calafiura, P., Childers, T., Fisk, I., Girone, M., Gutsche, O., Hufnagel, D., Klimentov, A., Lancon, E., Martinez Outschoorn, V. I., & Wuerthwein, F. (2025). Strategy For HPC Integration (US) Whitepaper. Zenodo. https://doi.org/10.5281/zenodo.17854525

While these individual initiatives have achieved considerable success within their respective domains, the increasing convergence of scientific requirements, as outlined in **Section 2** of this document, and the scale of emerging challenges necessitate a deeper integration between these large scale initiatives while better accounting for the requirements of the scientific users.. By having a coherent, well integrated and accessible scalable compute and data continuum[10]. The science community would benefit from economies of scale, increased efficiency, and the lowering of thresholds in difficulty for using the large-scale computing resources available in Europe. These challenges are not only about raw computing power. They also raise questions about how infrastructures should be designed and operated in a world of growing data volumes and increasingly complex workflows. Future architectural choices will need to build on the strengths of loosely coupled approaches while carefully assessing where strongly coupled models remain justified by scientific need. Systems must accommodate diverse computing architectures, different data types, and workflows that span facilities and domains. HPC, particularly at supercomputing centers, typically prioritizes tightly coupled, large-scale parallel jobs. As a result, operators are concerned that large ensembles of short, data-driven tasks could overload shared services (such as the Slurm scheduler database) or reduce overall system efficiency. In contrast, data-intensive communities tend to prioritize flexibility and high-throughput task execution.

Recent evaluations of the environmental impacts of the IT industry[11,12] have shown that these are rapidly increasing such that an absolute reduction is required to meet the Paris agreement[13] on climate change and, more generally, to stay within the planetary boundaries. Worldwide, IT is responsible for as much greenhouse gas (GHG) as double that of Canada or 5.5 times that of France, and its impact amounts to 40% of every connected individual's sustainable budget. As a leading European example, in France, it represents 11% of the total electricity consumption, 4.4% of the total GHG emissions distributed as 50% for terminals, 46% for datacentres and 4% for networks. These encompass significantly more than scientific IT infrastructures, with important contributions from the consumer market and the Cloud. The forthcoming increase in capacity must be carried out responsibly, that is sustainably, for the corresponding science to remain socially acceptable. This calls for environmental management of the compute and data continuum described in this blueprint. Life-cycle analysis, from design to disposal, is prescribed to provide a systemic view, allowing for balancing the impacts of successive phases and limiting the total footprint. For existing systems this will account for realised impacts and guide usage, maintenance and end-of-life policies, while for future ones, it will allow trade-offs between capacity increase and footprint reduction to make the growth sustainable.

This Technical Blueprint outlines a framework for a European Compute and Data Continuum that integrates computational resources, data management systems, and orchestration into a set of interdependent capabilities supporting large-scale scientific discovery. The framework is based on input from the **SPECTRUM Community of Practice (SPECTRUMCoP)**, analysis of use cases from HEP and RA, and a review of existing European e-infrastructures, as well as existing recommendations for their eco-design.

It is important to recognise that the communities addressed by SPECTRUM encompass distinct categories of users, each with different operational needs and interaction patterns. These include long-tail scientists who require accessible, lightweight interfaces and opportunistic access; software developers and pipeline engineers responsible for adapting and optimising complex scientific workflows; and instrument and observatory operations teams whose priorities centre on reliability, sustained throughput, and predictable long-term resource commitments. Explicitly differentiating these user groups helps align technical solutions with their respective expectations and ensures that the blueprint remains relevant across the full spectrum of scientific stakeholders.

---

[10] Within this document, the compute and data continuum is defined as a collaborative system of systems that links data centres, HPC facilities, cloud and quantum technologies, networking, and scientific instruments into an interoperable ecosystem. It relies on shared interfaces, authentication, orchestration and monitoring to support cross-facility workflows.

[11] Impacts environnementaux du numérique dans le monde, third edition, Association Green IT, 2025 (EENM 2025)

[12] Évaluation de l'impact environnemental du numérique en France, BRILLAND Thomas, FANGEAT Erwann, MEYER Julia, WELLHOFF Mathieu, ADEME 2025

[13] https://unfccc.int/process-and-meetings/the-paris-agreement

This document represents the culmination of extensive collaborative effort involving leading European scientific organizations and infrastructure providers. The resulting framework provides both tactical guidance for near-term infrastructure decisions and strategic vision for long-term capability development that will enable breakthrough scientific discoveries across multiple domains while establishing European leadership in scientific computing.

1.2. Methodology

The development of this Technical Blueprint employed a comprehensive methodology that integrated community engagement, technical analysis, and strategic planning. This approach ensures that the resulting recommendations reflect both current scientific requirements and future technological possibilities while remaining grounded in practical implementation considerations.

Community of Practice Engagement
The SPECTRUM Community of Practice (CoP) served as the primary vehicle for gathering input from the broader scientific community and infrastructure providers. Through a series of working groups focused on data management, workflow orchestration, computing environments, software tools, scientific use cases, and facilities management, the Community of Practice provided detailed feedback on current challenges and future requirements. The working groups conducted regular meetings throughout the project period, with participation from 65 experts representing diverse scientific domains and infrastructure providers across Europe.

A comprehensive survey administered to the CoP and the broader European scientific and computing communities captured quantitative data on resource utilization patterns, technical requirements, and policy preferences. The survey received around 75 substantive responses covering topics ranging from authentication and authorization preferences to energy efficiency concerns and emerging technology adoption plans. These responses, analyzed in the SPECTRUM Deliverable D3.1, provided empirical grounding for many of the technical recommendations contained in this blueprint.

Analysis Framework for Technical Use Cases
A detailed analysis of 14 representative use cases spanning HEP experimental programs, RA observing projects, and emerging AI/ML applications in both domains was carried out. Each use case analysis followed a standardized framework examining scientific challenges, storage requirements, data transport needs, computational demands, workflow management approaches, access patterns, and gap identification.

The use case methodology enabled systematic comparison across different scientific applications, identification of common technical requirements, and assessment of how current infrastructure capabilities align with projected needs. The use cases ranged from mature production workloads from HEP experiments and RA instruments to the development, analysis and workloads of small teams or individual scientists. To enable the use of larger and more complex (distributed) systems, particular attention focused on scaling challenges, heterogeneous computing requirements, data lifecycle management needs, and workflow orchestration complexity that characterizes next-generation scientific applications. The use cases are described in section 2.2.

Integration with previous SPECTRUM deliverables ensured consistency with the access policy (D5.2), infrastructure landscape assessment (D5.3), use cases (D5.1), and the SPECTRUM CoP (D3.1). This integration enabled comprehensive understanding of both technical capabilities and operational frameworks necessary for effective scientific computing support.

Analysis Framework for Infrastructures
The methodology included systematic assessment of 20 European e-infrastructures representing HPC, HTC, cloud, and data-oriented platforms. Assessment criteria encompassed technical capabilities, access policies, operational practices, and strategic development plans. This comprehensive coverage enabled identification of current capability gaps, potential integration opportunities, and strategic development priorities.

Infrastructure assessment results informed both technical architecture recommendations and policy framework development. The analysis revealed significant diversity in technical approaches, access models,

and operational practices that must be accommodated within federated infrastructure frameworks while enabling coherent user experiences and efficient resource utilization.

Community Feedback on the First Draft
The first version of this document was released for broader community feedback in December of 2025. Over the subsequent two months, more than 30 scientific computing and data experts from across Europe provided their feedback and suggestions. During the third week of February 2026, we also hosted an in-person event in Barcelona where experts from the community could attend in order to discuss and debate proposed changes to the Technical Blueprint and SRIDA. Overall, the feedback was very positive and constructive, and most of the suggestions received have been incorporated, resulting in the much improved version of the document presented here. We note that a major rewrite of Section 3 — "Sustainability" was undertaken post-consultation, including more actionable recommendations.

1.3. Related initiatives

Several European and international initiatives are tackling challenges that overlap with the objectives of SPECTRUM. While each has its own scope, together they provide important context and complementary efforts that inform the design of a European compute–data continuum.

In Appendix A, we give a brief description of a non-exhaustive selection of the following initiatives, in no particular order.

- JENA (Joint ECFA–NuPECC–APPEC) Activities
- InPEx (International Post-Exascale Project)
- ETP4HPC's Transcontinuum Initiative (TCI)
- EUCloudEdgeIoT and the OpenContinuum
- IPCEI-CIS (Important Projects of Common European Interest – Cloud Infrastructure and Services)
- The EuroHPC Federation Platform (EFP)

Together, these initiatives highlight a shared direction toward federated, interoperable, and sovereign digital infrastructures. JENA demonstrates cross-domain collaboration within science, InPEx offers a global research and policy perspective, TCI and OpenContinuum provide architectural blueprints for integration, IPCEI-CIS addresses sovereignty at industrial scale, and EFP tackles practical usability of EuroHPC systems. SPECTRUM's contribution is to bring the specific, large-scale needs of HEP and RA into this landscape, translating broad architectural visions into concrete requirements for scientific workflows, data volumes, and long-term sustainability. However, many of these capabilities are not yet technically implemented and where they are, they are often not interoperable with each other.

2. Preliminary work

The preliminary work carried out within SPECTRUM has provided the foundation for this Technical Blueprint. Through detailed analysis of representative use cases in HEP and RA, engagement with the Community of Practice, and alignment with existing European e-infrastructures, the project has identified key requirements, gaps, and opportunities for enabling a compute and data continuum. This work has not only clarified the technical challenges that must be addressed but has also highlighted commonalities across the two scientific domains, offering a pathway toward shared solutions for data intensive research. The findings presented here directly inform the design of the capability map (described in detail in Section 4), ensuring that the proposed continuum is grounded in real community needs. The subsections that follow describe the SEPCTRUMCoP (2.1), present representative use cases from HEP and RA (2.2), examine the current landscape and its gaps (2.3), and analyse the requirements for interoperable access policies (2.4).

2.1. The SPECTRUM Community of Practice

This section reports WP3 and WP4 activity describing the setup and outputs of the SPECTRUMCoP, the Knowledge Hub, the domain-wide survey and a first set of findings and directions that will feed into the Technical Blueprint and the Strategic Research, Innovation & Deployment Agenda (SRIDA).

As noted in the previous section, SPECTRUM WP3 established a Community of Practice to gather informed, cross-domain input (primarily HEP and RA, extended to related domains) on future compute and data needs for exabyte-scale science. The CoP is intended to remain an active, post-project forum to support alignment between research communities and e-Infrastructures and to provide the evidence base for SPECTRUM deliverables (Technical Blueprint, SRIDA). After the conclusion of WP3 in March 2025, WP4 took over the responsibility of the CoP.

2.1.1. The Knowledge Hub

The CoP created a Knowledge Hub, which is a Confluence-based repository and coordination space for WG outputs, meeting minutes, and curated community documents. It was chosen for familiarity and integration with the existing EGI Confluence site to take advantage of the existing platform.

The Hub collects reference documents and corresponding metadata while holding more than 60 curated entries spanning experiment roadmaps, technical reports, conference papers and policy inputs. Most activity contributing to the Hub has come from SPECTRUM members during synchronous meetings rather than asynchronous edits. The Hub is explicitly intended to inform the Technical Blueprint and SRIDA.

2.1.2. Findings

This subsection summarizes the findings from the knowledge hub and the survey presented in D3.1.

Career and skills
A recurring concern is the shortage of stable career paths for domain-specific software and data engineers; many respondents cited the risk that personnel will migrate to industry, threatening capacity to design and operate the large systems required in the coming decades. The report flags this as requiring high-level domain attention.

Software evolution
There is a consistent call to modernise software stacks to exploit heterogeneous architectures (multicore, vector units, GPUs, FPGAs) and to improve portability ,performance, and environmental sustainability. This includes emphasis on parallelisation, portability strategies and tooling.

Artificial Intelligence
AI/ML is regarded as a foundational technology to be integrated across software ecosystems (both as user tools and for optimisation), and the community expects growing dependence on AI over the next 5–10

years.

FAIR, open and cross-domain collaboration
The feedback from the survey emphasises adherence to FAIR principles (Findable, Accessible, Interoperable, and Reusable) for data and software, and recommends closer inter- and extra-domain collaboration (e.g., through shared repositories, standards and common toolchains) to avoid siloing.

Cloud, HPC and virtualization
The community recognises an increasing role for HPC computing as well as efficiently used scientific cloud resources. Surveyed e-Infrastructure providers report support for virtualization, and the CVMFS service is widely supported as a distribution method, a favourable unifying trend for deployment portability.

Network and node connectivity
A persistent technical constraint identified is network connectivity from compute nodes (external networking). While many centres indicate willingness/ability to enable needed connectivity and bandwidth, the report notes that promised capability must be validated against realistic use cases.

Resource allocation duration
Respondents highlighted the need for longer-term resource allocations to enable sustained workflows (rather than short, months-scale grants); the survey indicates that longer allocations are possible in many centres, but the report flags a policy and funding engagement need with infrastructure funders to secure such arrangements.

2.2. Representative use cases

This section provides a very brief overview of the SPECTRUM D5.1 report on "**Representative use cases: analysis and alignment**", which analyzes data-intensive scientific workflows from HEP, RA and elsewhere to inform the development of a European compute and data continuum for the Exascale era. The report analyzed thirteen representative use cases to identify current and future requirements in the areas of compute resources, data storage/processing capabilities, and essential services. The resulting analysis identifies cross-cutting requirements across the use cases as well as their potential impact on other areas of science.

For details on the selected use cases, please refer to SPECTRUM Deliverable D.5.1. Here we simply list the experiments and use cases that were studied, while recognizing that observatories like LOFAR2.0 and SKAO have a much broader science case than that captured by this analysis (see 'SKA1 Scientific Use Cases[14]). At the time of writing these use cases, many SRCNet architectural decisions were still pending, and the workflows for the scientific cases had not yet been developed. As such, we decided to focus on a small number of representative use cases that broadly encompass the kind of science that will be done with SKAO and LOFAR2.0.

High Energy Physics Use Cases

The four major WLCG/LHC experiments:

- CMS
- ATLAS
- LHCb
- ALICE

Selected AI Applications in HEP:

[14] Wagg et al.: SKA1 Scientific Use Cases (2021). https://www.skao.int/sites/default/files/documents/d35-SKA-TEL-SKO-0000015-04_Science_UseCases-signed.pdf/

- AI-based particle flow reconstruction[15,16,17,18]
- LLMs for the CERN accelerator complex[19]

There are numerous other AI- and ML-based algorithms and applications in HEP[20,21,22,23,24].

Radio Astronomy Use Cases

SKA Observatory (SKAO):

- SRCNet and 21 cm HI Fluctuations
- Star Formation History

Other Radio Astronomy Initiatives:

- LOFAR Extragalactic Surveys
- Fast Radio Bursts and Transients

Other Related Use Cases

- MISTRAL (Meteo Italian Supercomputing Portal)
- LIGATE Molecular Dynamics
- Simulation of plastic neural networks in the brain

2.2.1. Key Findings and Recommendations

The key recommendations and findings as reported in SPECTRUM Deliverable D5.1 are as follows:

1. Unprecedented Data Growth: Both HEP and RA communities are entering an era of steep growth in data generation that will exceed current infrastructure capabilities.
2. Heterogeneous Computing Evolution: there is a clear trend toward mixed CPU/GPU environments, with a growing importance of specialized AI hardware.
3. Complex Workflow Requirements: Advanced workflow management systems are needed to orchestrate multi-step processes across distributed resources.
4. Data Federation Essentials: Robust data federation mechanisms are required to support efficient discovery, access, and transfer across multiple sites.
5. Resource Planning Horizons: Multi-year allocations for storage and compute are needed to enable better integration of HPC resources into long-term scientific planning.

---

[15] Pata, J., Duarte, J., Vlimant, JR. et al. MLPF: efficient machine-learned particle-flow reconstruction using graph neural networks. Eur. Phys. J. C 81, 381 (2021). https://doi.org/10.1140/epjc/s10052-021-09158-w

[16] Machine Learning for Particle Flow Reconstruction at CMS, Joosep Pata et al 2023 J. Phys.: Conf. Ser. 2438 012100 (2023), doi.org/10.1088/1742-6596/2438/1/012100

[17] Pata, J., Wulff, E., Mokhtar, F. et al. Improved particle-flow event reconstruction with scalable neural networks for current and future particle detectors. Commun Phys 7, 124 (2024). https://doi.org/10.1038/s42005-024-01599-5

[18] Fine-tuning machine-learned particle-flow reconstruction for new detector geometries in future colliders Farouk Mokhtar et al. Phys. Rev. D 111, 092015 (2025) https://doi.org/10.1103/PhysRevD.111.092015

[19] AccGPT: A CERN Knowledge Retrieval Chatbot, Florian Rehm et al. EPJ Web of Conferences 337, 01279 (2025) https://doi.org/10.1051/epjconf/202533701279

[20] A Living Review of Machine Learning for High Energy Physics, https://iml-wg.github.io/HEPML-LivingReview/

[21] Machine Learning in High Energy Physics Community White Paper, Kim Albertsson et al 2018 J. Phys.: Conf. Ser. 1085 022008, doi.org/0.1088/1742-6596/1085/2/022008

[22] R.L. Workman et al. (Particle Data Group), Prog. Theor. Exp. Phys. 2022, 083C01 (2022) and 2023 update, https://pdg.lbl.gov/2023/reviews/rpp2023-rev-machine-learning.pdf

[23] Review of Machine Learning for Real-Time Analysis at the Large Hadron Collider experiments ALICE, ATLAS, CMS and LHCb, Laura Boggia et al. https://doi.org/10.48550/arXiv.2506.14578

[24] Mondal, S., Mastrolorenzo, L. Machine learning in high energy physics: a review of heavy-flavor jet tagging at the LHC. Eur. Phys. J. Spec. Top. 233, 2657–2686 (2024). https://doi.org/10.1140/epjs/s11734-024-01234-y

6. Standardization Needs: Common facility standards and implementation of these standards would reduce costs and complexity of future integrations.
7. Authentication Evolution: Authentication and authorization systems must be federated and evolve to balance security requirements with the need for automation.

The report concludes that addressing these challenges is essential for maintaining Europe's position at the forefront of scientific discovery. The SPECTRUM project aims to guide this evolution by fostering an ecosystem that enables groundbreaking scientific discoveries in the Exascale era.

The findings demonstrate that next-generation scientific infrastructure must support diverse computing paradigms simultaneously, with flexible architectures capable of handling heterogeneous workloads. Data management systems must scale to exabyte levels while providing seamless access across distributed environments. Perhaps most importantly, closer collaboration between scientific communities can lead to more efficient resource utilization and accelerated innovation in shared tools and methodologies.

2.3. Landscape, Requirements and Gap Analysis

This section considers the SPECTRUM Deliverable 5.3 report on "**Landscape of RIs: technologies, services, gaps**", which surveys the technical characteristics of 20 European e-Infrastructures supporting research communities and data-intensive applications with significant computing requirements, particularly with HEP and RA use cases. D5.3 identifies gaps and proposes recommendations for the future evolution of Europe's e-Infrastructure landscape and is aligned with the recommendations provided in the companion SPECTRUM Deliverable D5.2 (Interoperable Access Policies: Analysis and Recommendations).

2.3.1. European e-infrastructure landscape

The infrastructures studied span HPC, High Throughput Computing (HTC), dedicated Data-oriented infrastructures, and Cloud e-Infrastructures.

- **HPC-oriented infrastructures**
  The EuroHPC Joint Undertaking (JU) forms the backbone of this landscape, comprising operational petascale systems[25] such as Deucalion, Discoverer, Karolina, Leonardo, LUMI, MareNostrum 5, Meluxina and Vega, with new systems JUPITER (Germany), Arrhenius (Sweden), Daedalus (Greece) and Alice Recoque (France) planned for 2025-26. National centres such as ICSC (Italy), GCS/JSC/HLRS/LRZ/NHR Alliance (Germany), RES (Spain), CSCS (Switzerland), EPCC (UK) and GENCI (France) complement these resources and will support future HEP and RA experiments.
- **HTC-oriented infrastructures**
  The WLCG serves the global HEP community with about 1.4 million CPU cores and 1.5 Exabytes of storage (2025). WLCG Tier-1 centers, the EGI HTC-oriented infrastructure, provide access for diverse scientific communities, together with upcoming HTC platforms that include regional centres such as the SRCNet.
- **Data-oriented infrastructures**
  They include the WLCG's data lake (across the world and with CERN in Switzerland as the central hub), the SRCNet, the EBRAINS neuroscience platform, the LOFAR Long-Term Archive (LTA), German's ErUM4 Data Hub, the ICSC National Center in Italy, the PUNCH4NFDI initiative and the Copernicus Earth-observation data space.
- **Cloud-oriented infrastructures**
  In contrast to the three categories of infrastructures above, Cloud-oriented e-Infrastructures often provide a combination of compute (HTC and increasingly HPC) and data services. This category includes for example: the EGI FedCloud, which offers a federated cloud interface that unifies resources across Europe; SURF's Data Processing Services; the Grid and Spider, are HTC platforms deployed using cloud services,, the European Open Science Cloud (EOSC) Federation, which federates national nodes and thematic pilots (the first 13 pilot nodes were announced in

[25] https://www.eurohpc-ju.europa.eu/supercomputers/our-supercomputers_en

March 2025[26]) and the Simpl Data Federation Platform, built under a commercial contract for the European Commission, which aims at unifying access to diverse European data spaces.

2.3.2. Proposed requirements

This section derives technical and operational requirements for future e-Infrastructures. The requirements for use cases presented in Deliverable 5.1 (Representative use cases: analysis and alignment) and described above, have been analyzed with the results included in this section. To ensure consistency, the study defined a template covering five key aspects:

- **Storage/Data Capabilities and Services,** detailed in section 2.3.2.1.
- **Data Transfer and Federation,** detailed in section 2.3.2.2.
- **Compute Capabilities and Services**, detailed in section 2.3.2.3.
- **Software Services, Interfaces and Stacks,** detailed in section 2.3.2.4.
- **Compute Federation Services,** detailed in section 2.3.2.5.

2.3.2.1. Storage/Data Capabilities and Services

Requirement #1: FAIR storage for observation/experiment data across domains
Across data-intensive scientific domains, agree on common interfaces and core implementation for e-Infrastructures that preserve observation/experiment data (including meta- and provenance data) and make it available to scientists and the public according to the FAIR principles.

The key motivation for such an architecture and implemented core components is to preserve the vast and rapidly growing amounts of observation/experiment data together with the relevant metadata and provenance information according to the FAIR principles. The capability to search for and locate data based on provenance and contents, and the ability to interoperate seamlessly with HTC and HPC infrastructures, will be other important requirements. The need for clear metadata is similarly important for facilities conducting photon and neutron science where the research communities are composed of many small groups dealing with increasingly large volumes of data.

To partially address the FAIR principles, although ambitious, it could be important to design an ontology that provides controlled vocabularies for the continuum. This ontology can include both high-level metadata and detailed descriptions of data and processes for software, hardware, and workflows. Ontologies can serve as a basis for developing uniform APIs across all operational fields.

Further Considerations (Requirement #1)

- Long-term storage of irreplaceable observation/experiment data is a key requirement across HEP, RA and many other science domains - efforts like WLCG, LOFAR LTA and SRCNet demonstrate how this can be done.
- The number of data-intensive research communities and the data volume is increasing sharply, with issues to get access to needed storage volumes.
- Common interfaces or core implementation for storing/providing experiment/observation data and provenance/metadata could leverage economies of scale and avoid duplication of work.
- Efficient interfaces to HPC and/or HTC infrastructures needs to be incorporated to enable seamless work across heterogeneous systems.
- The hard requirement to support international, non-EU researchers raises policy and technical issues.

Requirement #2: Appropriate provision must be made for data covered by the GDPR or other restrictions
All the scientific domains considered require some forms of data security and traceability, which become extremely important for life sciences and medicine, which require storing and processing of personally identifiable data in an e-Infrastructure. Thus, to optimise and standardise the technical implementation to support EC (and national) regulations (e.g., GDPR) effectively for different scientific use cases needs to become a central requirement of any future infrastructure.

[26] https://www.eosc-beyond.eu/pilots

Considerations (Requirement #2)

- Storing and processing sensitive (in particular special category) data or commercially valuable/sensitive data requires stringent data protection (policies and technical means). This includes the handling of user account data irrespective of scientific domain.
- Technical measures include encryption of data at rest/in transit/in memory, secure & fine-grained management of keys and access rights, extensive logging, vetting/validation of applications and audit capabilities.
- Care has to be taken when integrating support for storing and processing personal and in particular special category data with an open science based FAIR data storage/provisioning system; the strict protection rules and limitations for such data should not be applied to fully open scientific data,
- Potentially, extra security layers may be added (e.g. secure sandboxes). Viability of standard tools for data anonymisation or pseudonymisation should be investigated.

2.3.2.2. Data Transfer and Federation

Requirement #3: Automated, efficient data movement and data staging

Data movement and data staging between storage solutions within a single site, between sites in an e-Infrastructure, or between federated e-Infrastructures (such as between data storage and HTC/HPC systems) should be handled automatically and rapidly.

Further Considerations (Requirement #3)

- Data staging/transfer is often performed in job scripts/workflow steps using low-level and centre-provided/specific tools, complicating coding, maintenance and use.
- Highly efficient, multi-stream transfer methods are not universally adopted by e-Infrastructures yet.
- Standard interfaces to express data transfer to be automatically mapped to specific tools/protocols would address this - an example is **FTS** as used by **WLCG/EGI**, and SRCNet.
- Large transfers need to be scheduled externally from the sending and receiving sites. The capability for *third party transfers* is essential in large scale data management.
- Automated data transfer into/out of HPC centers will impact the security architecture and configuration of (HPC) centres.
- It may be that using open data transfer mechanisms would be optimal to proprietary or closed transfer protocols.

Requirement #4: Plan data transfer capacity according to research community requirements

For future computing e-Infrastructures, the evolving needs of research communities should drive the planning of data transfer and ephemeral storage capacities. Compute and long-term storage are usually planned more explicitly, while transient data movement and short-lived storage needs tend to scale unpredictably with workflow complexity, making them more sensitive to community-driven changes.

Further Considerations (Requirement #4)

- Capacity, latency, and performance of data transfer into/out of HPC/HTC centers must fit the demand.
- Data transfer requirements will likely evolve across the lifetime of expensive HPC, HTC, cloud and data driven resources; hence, the centers must be prepared for evolving their initial policies according to user needs.

2.3.2.3. Compute Capabilities and Services

Requirement #5: Enable portability of compute tasks across accelerated compute platforms

Enable compute tasks used by research communities to execute across different families/vendors of accelerated compute resources, helping to avoid unnecessary transfers and duplication when switching between vendors and improving overall efficiency.

Further Considerations (Requirement #5)

- Accelerator (mainly GPU) capacity has been vastly increased by HPC infrastructures.
- Portability of codes between different hardware accelerators is impacted by the proprietary nature of the CUDA software ecosystem.
- Portability to other classes of accelerators (such as FPGAs) is difficult for fundamental reasons; for example, FPGAs use spatial parallelism, have no fixed instruction set, rely on deeply customized

memory layouts, and require hardware-level design flows. These traits make automatic portability from GPU/CPU code inherently difficult.
- Portable programming models or portability layers for GPUs do exist, examples include Kokkos, Alpaka, SYCL, and oneAPI, and high-level AI environments, such as PyTorch and TensorFlow, show how backend optimisations can be made transparent to end-users.
- Software engineering tools should enable the rapid and efficient development and deployment of new services on cloud/HPC/edge architectures, taking into account non-functional constraints such as energy consumption and source code management. Such tools are too often tied to hardware solution providers and are therefore not interoperable. The slow pace of development of these tools means that developer productivity is progressing very slowly compared to the complexity of applications and hardware computing systems. The effort to provide such open tools is currently completely negligible in view of the challenges involved.

2.3.2.4. Software Services, Interfaces, and Stacks

Requirement #6: Establish common workflow systems supported by all e-Infrastructures
Identify a number of workflow systems covering the needs of key research communities and ensure support by e-Infrastructures.

Further Considerations (Requirement #6)
- There is no shortage of different workflow systems (for encoding workflows and orchestrating them), including general (e.g., PanDA), and domain-specific frameworks.
- HPC and HTC centers must provide interfaces to support orchestration of workflows.
- Supporting multiple workflow systems for different communities creates overhead and duplicated work (one example is the WLCG HTCondor-based workflows versus LEXIS for the EuroHPC JU federation platform).
- Recognizing the inherent challenges and potential impact on future innovations, improved collaboration and co-design between providers and scientific communities could reduce the set of workflow systems in a first step, and then ideally agree on a small number (ideally one) of systems across domains, and the time saved through such convergence could be re-invested in optimising how workflows map onto the infrastructure.
- External API access to the internal batch system in HPC centers would allow decoupling the workflow system from the one used in the fabric (typically Slurm).

Requirement #7: Establish common software stacks for compute applications
Establish common software stacks for research communities deployed on the relevant e-Infrastructures.

Further Considerations (Requirement #7)
- Similar to requirement #8 in D5.2, common software interfaces across different HW architectures and system configurations would simplify the life of end-users.
- Here, the focus is more on how to manage virtualisation techniques (e.g., containers) to provide specific, optimised stacks for combinations of CPUs, interconnect networks and accelerators (e.g., GPUs).

2.3.2.5. Compute Federation Services

Many of the use cases studied in detail in D5.1 explicitly require federation of compute resources. The EuroHPC JU Federation Platform is introducing an AARC-BPA-compliant AAI, enabling the use of institutional identities via eduGAIN across EuroHPC systems, along with interoperable (or federated) services for cross-site workflows and simplified, common interfaces for the use of the HPC systems, following previous initiatives like FENIX and Grid Computing platforms. Other HPC e-Infrastructures, such as CSCS and GENCI, provide similar federation capabilities between their constituent sites/systems. It should be mentioned that none of these support end user access privileges/quotas across all participating systems, and that automatic routing of accesses/compute tasks to the best suited systems is not available.

As mentioned in D5.3, WLCG federates its HTC resources for workflow execution; this is currently done via experiment specific layers, with centers providing an access point to internal resources (usually mediated by a batch system). In the case of SKAO, SRCNet will also operate as a federated system, but has been designed so as to be agnostic to the data analysis use case. The EGI federation provides a brokering service for scientific end-users and makes access to public and commercial providers of OpenStack Cloud

resources (EGI Federated Cloud) federated via a common AAI, accounting and image repository. Simpl (Smart Middleware Platform) is a secure middleware software platform, procured by the EC, that supports federated data access and interoperability in European sectoral data spaces which is still in the early stages of development.

- Onboarding of users by the governance authority.
- Catalogues for infrastructure & data, usage policies, and quality rules.
- UI and API for adding to the catalogues and for validating entries.
- UI and API for basic and extended searches within a single data space.
- Resource selection, request to use that resource, and establish a basic contract between the data provider and the consumer.
- Infrastructure deployment, data set access, and initial types of data processing.
- Secure communication between Simpl agents, logging of basic metrics for all user actions.

2.4. Interoperable access policies

This section considers the SPECTRUM D5.2 report on "**Interoperable access policies: analysis and recommendations**", which presents the results of an in-depth study on existing and proposed access policies for European e-Infrastructures, supporting research communities and data-intensive applications with substantial computing needs.

2.4.1. Access Policies, Gaps and Recommendations

This section presents the analysis results for the set of selected e-Infrastructures listed above. The content is derived from information made available online via the Web presences or service portals maintained by the e-Infrastructures covered, as well as published in papers and presentations, or obtained through direct communication with e-Infrastructure providers. The access policies analysis and the methods are based on the following common template:

- **Obtaining Access**
- **Access Tracks and Modalities**
- **Summary of Accessible Resources**
- **Access Management and Security**
- **Rules and Assurances / Monitoring, Evaluation, and Evolution**

Obtaining Access
From the analysis of access policies, two distinct methods of handling access requests and providing access emerge:

- Allocating resources according to the result of reviewing a specific access proposal according to the rules of the resource provider or of the organisation responsible for a resource call-for-proposal, which focus on scientific merit and novelty, and often require a peer review. Access is granted for a specified period (usually one year or less, in some cases two years), and extension proposals are typically supported by most providers. Otherwise, the mechanism is designed to disallow repetitive access proposals, and in some cases, it also restricts principal investigators (PIs) from submitting multiple different access proposals.
- Based on the agreement between a research community and an e-Infrastructure, and in some cases involving a third party as a broker, access is provided to scientists based on their demonstrated membership in that research community. The allocation period can be an arbitrary amount of time, and entitlement lapses when the scientist leaves the research community (or "virtual organisation").

Gap #1: Long-term assured access and planning of resource allocations:

- Grant-based access is limited in time (often 1–2 years or less).
- Grants may be extended, but this is not guaranteed and hence problematic for long term planning.
- While perfectly adequate for some use cases, other cases need guaranteed access for longer time periods (e.g. 3–5 years), compatible with the long range planning and funding cycles appropriate for these experiments.

Recommendation #1: In addition to short-term access paths, adopt longer-term, flexible resource allocation processes.

Gap #2: Resource allocations that are valid across a single e-Infrastructure:

- In many cases (such as EuroHPC JU), allocations are good for specific hosting sites/systems only.
- CoP members have expressed a strong desire to use resources seamlessly across an e-Infrastructure to improve how their workloads are executed. In particular, they seek to take advantage of burst capacity—temporarily scaling beyond local limits by accessing external resources during peak demand—while also optimizing efficiency for specific workload components and strengthening resilience against failures or disruptions.
- *Recommendation #2: Enable e-Infrastructure-wide use of resource allocations and quotas.*

Access Tracks and Modalities

Traditionally, HPC and HTC centers focus on batch processing, governed by a scheduling/orchestration system like Slurm. Here, end-users submit job scripts (containing the amount and type of compute resources required), which are run at a time determined by the scheduling system in unattended mode. This approach enables the efficient use of available HPC resources and is well-suited to large-scale parallel jobs. Interactive access is possible in two ways: via special head or login nodes (potentially competing with other users) or sets of (partial) nodes allocated via the batch system (with exclusive access).

Gap #3: Interactive access to significant-scale computing resources:

- Problems include delays in availability (due to scheduling), and potential underuse of resources (in batch scenarios).
- Compromise to be reached between quick availability and performance guarantees.

Recommendation #3: Extend scheduling/orchestration to support on-demand interactive compute use cases with access to sizable nodes.

Gap #4: High-level end-user interfaces:

- Most e-Infrastructures do not offer high-level, abstracted interfaces for common application/workloads; for example, like LOFAR pipeline portals or HEP systems like PanDA, which let users run complex workflows without dealing with scheduler flags, modules, or environment setup.
- Domain scientists should not need to interact with system details and brittle application interfaces, such as low-level scheduler options, module setups, environment conflicts, or fragile command-line tools.

Recommendation #4: Introduce high-level general and domain-specific user interfaces, valid across a multitude of sites.

Access Management and Security

Common mechanisms used by the federated e-Infrastructures studied here include using password-protected certificates/keys (such as SSL key pairs) provided by the end-user, or relying on tokens or time-limited keys provided by a central single-sign-on service (which in turn uses userid/password or certificate/key authentication). Due to security concerns, centers have started to enforce multi-factor authentication (MFA), which involves an "interactive" authentication step involving a resource (such as a mobile phone) in the possession of the end user and the end user him/herself. For the widely adopted authentication method using SSH key pairs, end-users can avoid repeatedly entering the private key password, for instance, by keeping an initial connection open for a full session or by automating the password entry using Linux mechanisms. Unfortunately, this creates security vulnerabilities, particularly for the second example, which would require the storage of cleartext passwords on file or in memory.

Gap #5: Adoption and lack of federation of end-user identities (also across e-Infrastructures):

- Federated e-Infrastructures abstract away from local user identities, allowing users to access distributed resources through a unified identity rather than relying on site-specific accounts.
- As motivated in SPECTRUM D5.2, end-users working across different federated e-Infrastructures would benefit from common identities/AAI systems across the different infrastructures[27].

Recommendation #5: Introduce common and interoperable AAI services across European e-Infrastructures with support for service accounts.

[27] For example, proposed by AARC – https://aarc-community.org/aarc-tree-project/.

Gap #6: Unattended execution of long-running workflows and services:

- Increasing reliance on MFA-based authentication methods, which do not support the scenario of unattended workflows running for days/weeks, introduce the need for re-authentication (manual or automatic).
- Access privileges handed out after interactive or MFA authentication have a limited period of validity (often one day).
- After that period, workflow steps would need to interactively re-authenticate.

Recommendation #6: Ensure Reliable and Unattended execution of long-running workflows.

Rules and Assurances / Monitoring, Evaluation, and Evolution

All considered e-Infrastructures require their end users to comply with acceptable use policies; while they differ in detail across the infrastructures, they contain a core of rules banning malicious behaviour (which would impact the operation and/or other end users), oversubscription of scarce resources (such as access or login nodes, or flooding a batch system with jobs), and use for commercial purposes (unless explicitly approved). In general the e-Infrastructures give "best effort" assurances regarding availability and operation of their resources, and don't accept liability for direct or consequential damages caused, for instance, by non-availability of services, technical or operational faults, or activities of their personnel.

System and center operators all monitor and evaluate the performance of their local infrastructure in site-specific ways. This includes uptime, incidence of faults, utilization of compute, storage, and network resources, as well as usage statistics derived from end-user and project accounting. This data informs the operation of resources/sites (such as the scheduling of maintenance or the incremental addition of resources), and it forms the basis for justifying public funding received from funding authorities or projects/communities. The key observation is that such measures are handled in a site-specific (and often non-public) way for most of the studied e-Infrastructures.

Gap #7: Efficient provision of standard, low-level software interfaces across different (accelerator) systems:

- software environments differ considerably between e-Infrastructures (often even between different sites), in particular if performance/efficiency is a prime concern.
- Standardised software stacks with scalable software distribution and efficient build systems have been proposed / are available (such as Easybuild and Spack), in addition to containers.
- Need widespread agreement on a set of commonly provided low-level software interfaces for application portability and effective ways to support different (accelerator) systems.

Recommendation #7: Provision efficient standard, low-level software interfaces.

Gap #8: Organised feedback/improvement/planning loop:

- Organised and effective collaboration on co-designing and improving e-Infrastructures is happening in places (such as **WLCG** and **SRCNet**), yet is not common.
- Rolling such schemes out between all relevant end-user communities and e-Infrastructures would improve planning and help to ensure good support of end-users.

Recommendation #8: Establish feedback/improvement/planning loops for research communities and e-Infrastructures.

2.5. Consolidated Gaps/Requirements and Recommendations

This section summarizes the gaps, requirements and recommendations found in sections 2.1 to 2.4. Some of the gaps and recommendations from the previous sections are similar and have been merged for clarity. The fact that similar gaps have been identified in different places show that these requirements are grounded in the real needs of the HEP and RA communities and that these communities share many challenges and opportunities.

Table 1 summarizes and lists the gaps/requirements identified during the first phase of the SPECTRUM project and briefly describes the project consortium's recommendation. It also indicates from where the gap/requirement was identified by pointing to the Subsection in Section 2 where it was covered. Finally it also indicates which capabilities in the Compute and Data Continuum capability map that are involved in

implementing each recommendation. The full list of capabilities that we propose should be part of the future data continuum are presented in the capability map presented in Section 4.

Table 1: Consolidated gaps/requirements and recommendations from the study of the landscape of RIs, use cases from HEP, RA and beyond, as well as interoperable access policies.

#	Gap/Requirement	Recommendation	Source Section	Capability
1	**Long-term resource provisioning** Current grant-based allocations are short (1–2 years) and fragmented by site; HEP and RA need assured access on decadal timescales.	As a complement to existing short-term funding programs, we recommend establishing longer-term (e.g. 3–5 year) flexible allocation frameworks, federated across infrastructures, with transparent accounting and monitoring.	2.1, 2.2, 2.3, 2.4	Resource Federation, Operational Sustainability
2	**Federated access and identity management** End-users face fragmented identity systems across e-infrastructures, complicating multi-site workflows. Authentication often interrupts unattended, long-running jobs.	Deploy interoperable AAI frameworks across European e-infrastructures, supporting single sign-on, attribute-based authorization, and persistent tokens for long workflows.	2.2, 2.4	Security & Trust, Resource Federation
3	**Interoperable allocation and accounting** Allocations and quotas are bound to specific e-infrastructures, limiting flexibility. Accounting practices differ across infrastructures.	Enable e-infrastructure-wide allocations and harmonized accounting, so projects can burst across systems and providers while ensuring fair contributions and usage transparency.	2.2, 2.3, 2.4	Resource Federation, Operational Sustainability
4	**High-level and portable software environments** Software stacks differ widely between sites; portability across CPUs, GPUs, and other accelerators is inconsistent. Scientists often face brittle interfaces.	Provide standardized, efficient software stacks (containers, CVMFS) and high-level domain-specific interfaces, enabling reproducibility and accelerator portability.	2.1, 2.2, 2.3	Software & Execution, Compute Continuum
5	**Workflow execution and orchestration** Workflow engines are fragmented across communities (e.g. HTCondor in HEP vs. LEXIS in EuroHPC). Cross-facility workflows lack resilience and provenance capture.	Co-design and converge on a set of common APIs for workflow systems, with support for unattended execution, provenance tracking, and cross-infrastructure orchestration.	2.2, 2.3	Orchestration & Workflows, Software & Execution
6	**Data federation and FAIR principles** No unified approach to FAIR storage, metadata, and provenance across domains	Establish interoperable federated data management with FAIR-compliant storage, metadata catalogs, and efficient interfaces to HPC/HTC, ensuring long-term access and reproducibility.	2.1, 2.2, 2,3	Data Continuum, Operational Sustainability

#	Gap/Requirement	Recommendation	Source Section	Capability
7	**Data transfer and staging** Data movement across sites often relies on ad-hoc scripts and site-specific tools; transfer capacity planning is uneven.	Automate large-scale, operated by other parties, efficient data movement with standardized protocols, and plan data transfer/storage capacity jointly with community needs.	2.2, 2.3	Data Continuum, Compute Continuum
8	**Real-time and low-latency processing** A small number of RA use cases (e.g. transient detection) and some HEP triggers require near-real-time data analysis.	Extend compute continuum capabilities with real-time, low-latency processing.	2.3	Compute Continuum, Real-time & Low-Latency Computing
9	**Security, privacy, and sensitive data handling** Personal or commercially sensitive data requires stronger safeguards; GDPR compliance adds complexity.	Provide flexible security architectures: robust encryption, key management, auditing, anonymization/pseudo-anymization, and possibly separate high-protection infrastructure paths.	2.2, 2.4	Security & Trust, Data Continuum
10	**Monitoring and observability** Current infrastructures lack consistent, cross-site monitoring of workflows, performance, and energy usage.	Deploy monitoring and observability frameworks that capture standardized metrics describing infrastructure health, application performance, provenance, and sustainability.	2.3	Monitoring & Observability, Operational Sustainability
11	**AI/ML integration** AI/ML workloads are increasingly central and are being investigated in every step of the data-processing pipelines, from data-acquisition to simulation and analysis.	Enhance services in existing academic and research AI/ML platforms to support distributed training, inference deployment, and access to foundation models, integrated with HPC and domain workflows.	2.1 2.2, 2.3	Advanced Analytics & AI/ML, Software Distribution & Execution
12	**Co-design and feedback loops** Systematic co-design between science communities and infrastructure providers is patchy; planning often mismatches user needs.	Establish structured feedback loops (like WLCG/SRCNet) across domains, embedding co-design in infrastructure roadmaps, procurement, and evaluation.	2.1 2.2, 2.3, 2.4	Operational Sustainability, Resource Federation
13	**Workforce sustainability and development** Long-term sustainability depends on stable funding, governance, and skilled personnel. Current models are fragmented.	Develop coordinated European funding/governance models, align with national contributions, and invest in training and career pathways for research software engineers and data stewards.	2.1 2.2, 2.3	Operational Sustainability
14	**Interactive access to large-scale computing** Interactive access to small/medium partitions is possible. Problems	Extend scheduling/orchestration to support interactive compute use cases (including those that have significant resource needs).	2.4	Compute Continuum, Orchestration

#	Gap/Requirement	Recommendation	Source Section	Capability
	include delays in availability, and potential underuse of resources.	Compromise to be reached between quick availability, performance guarantees and efficient use of resources.		and Workflows
15	**High-level end-user interfaces** Most e-Infrastructures do not offer high-level, abstracted interfaces, meaning domain scientists are often required to understand system-level details and interact with inflexible or fragile application interfaces.	Introduce high-level general and domain-specific user interfaces. (like workflow portals, Jupyter, or CARTA).	2.4	Software Distribution & Execution, Orchestration & Workflows
16	**Unattended execution of long-running workflows and services** Two-factor authentication can impact execution of long-running workflows and services. Access established after interactive or MFA authentication has a limited period of validity. Workflow steps need to interactively re-authenticate.	Ensure reliable and unattended execution of long-running workflows and user-operated services, without the need for frequent (manual) re-authentication.	2.4	Orchestration & Workflows, Security & Trust

3. Sustainability

Sustainability is a core concern for large research infrastructures such as those in HEP and RA, given their long lifetimes and the need for sustained contributions over decades. Ensuring that science remains reproducible requires that intellectual, technical, and financial investments build cumulatively on existing knowledge for future generations. In IT terms, this means preserving data and software to enable reuse and support open science, while also maintaining computing and storage infrastructures in an environmentally sustainable way so they can operate over long periods of time with acceptable impact.

Achieving this also depends on cross-cutting qualities such as efficiency, quality, and flexibility, which help deliver results while minimizing the use of resources like energy, water, and hardware. A complementary principle is sufficiency—using only what is necessary to meet scientific goals—which is essential for reducing overall environmental impact. Realizing this in practice relies on skilled teams with a deep understanding of the systems they manage, built through experience and supported by a stable, sustainable workforce.

3.1. Data and software preservation and reuse

Significant efforts are devoted to building data and software archives that not only collect assets but also ensure their continued usability despite the evolving IT landscape. However, their use is often limited to individual communities or subgroups, contributing to the current reproducibility challenges in science[28]. Enabling broader reuse of data and software would strengthen scientific results, deepen data interpretation through successive analyses, and improve software robustness via independent validation. These goals underpin FAIR data management and Open Science, supporting sustainability by increasing trust in scientific outputs, reducing duplication of effort, and enabling incremental software development rather than repeated reinvention.

Extending reuse across communities would amplify these benefits by fostering interdisciplinarity and sharing the substantial intellectual and technical investments behind data analysis, thereby allowing it to scale. In the Exabyte era, the compute and data continuum should promote software reuse to reduce the span of software used in production (i.e. its heterogeneity). This would facilitate development and efficiency on the one hand, as well as improve deployment on the other hand by relying on a concentrated set of software which are optimised for the platform, compute-intensive tasks and I/O. Technically speaking, facilitating the reuse of existing software requires flexible bindings in multiple programming languages as well as assisting developers in identifying candidate existing software. The latter could be achieved through libraries, software repositories, and indexing software repositories to match their contents to a set of functional needs. One can also generate wrappers for using existing software from another language using an API. In terms of policy, beyond the production of software for a given purpose, projects should be encouraged to produce or contribute to libraries in order to both increase reusability and harmonize the software stack.

3.2. Environmental sustainability

This section addresses environmental sustainability from the perspectives of both scientists running software on the infrastructure and the IT personnel responsible for operations. While neither group directly controls the fabrication or end-of-life phases of equipment, they can influence these upstream and downstream impacts through procurement and disposal decisions as key actors in the value chain.

Recommendations in this section go beyond improving the energy efficiency of systems, as such improvements have historically not led to lasting reductions in environmental impact. Cheaper and easier access to resources has tended to increase demand—because it is highly price-elastic—so that overall usage growth outweighs efficiency gains. This phenomenon, known as the rebound effect or Jevons' paradox, is difficult to counter because it is driven by future and often unanticipated uses.

[28] Marshall-Cook, Joanna and Martin Farley. "The hidden sustainability cost of the reproducibility crisis." *Nature Reviews Physics* 6 (2023): 4 - 5. https://doi.org/10.1038/s42254-023-00674-0

For this reason, we emphasise extensive system properties[29] that scale with size and relate them directly to absolute measures—namely, environmental impacts. This approach places resource use in context, recognising it as the result of trade-offs between direct and indirect environmental impacts and societal benefits. Energy use, for example, is not inherently negative, but it should be balanced against the benefits it delivers and the impacts it generates.

The focus is often on energy, largely for cost or procurement reasons. We consider instead that sustainable decision-making should rely on Life-Cycle Assessments (LCA), based on ISO's 14040 and 14044 norms, and encompass all 16 impact categories, as defined in the Product Environmental Footprint method recommended by the European Commission[30], in order to avoid transfers between phases or across impacts, and allow for systemic reasoning. **Table 2** illustrates the result of such an analysis for a datacenter located in the University of Padova in Italy (with simplified impact categories). Depending on the category, the scale can be global (e.g. climate change) or local (e.g. water use). Spread out temporarily and geographically (as a result of the globalised fabrication and disposal) impacts are intrinsically difficult to manage. Additionally, the correspondence between impacts and planetary boundaries[31] is not straightforward due to the complexity of the processes at stake[32], which further calls for an open, science-based decision-making approach.

Lifecycle assessment for IT equipment in Europe shows that, as a result mainly of its short lifetime and the existing energy mix, the climate change impacts of the fabrication phase outweigh those related to usage. Replacing hardware to benefit from efficiency improvements is never energetically beneficial in such conditions[33]. Additionally, this does not account for other impacts, like acidification, eutrophication, depletion of abiotic resources, human and eco-toxicity which are mostly related to the materials used. Hence a top-level priority which flows down to the recommendations below is to use the hardware for as long as possible.

The power behaviours of IT devices are complex and highly non linear: fixed costs related to the power supply, standby and leakage powers are significant and CPU cores, for instance, are far from independent power-wise with behaviours being history-dependent. Modelling this is still largely a research topic but it is acknowledged that the most power-effective use of nodes is achieved when the load is high and that an idle one still consumes about half of its peak power. While this can impede absolute job performance because of resource saturation, concentrating jobs on a minimum number of resources, up to a certain point, allows first for fabricating fewer and then for more effective power management.

[29] As opposed to intensive properties like efficiency, which being independent of size, do not allow for bounding environmental impacts.

[30] Damiani, M., Ferrara, N., Ardente, F. "Understanding Product Environmental Footprint and Organisation Environmental Footprint methods." *JRC technical report 2022*. https://publications.jrc.ec.europa.eu/repository/bitstream/JRC129907/JRC129907_01.pdf

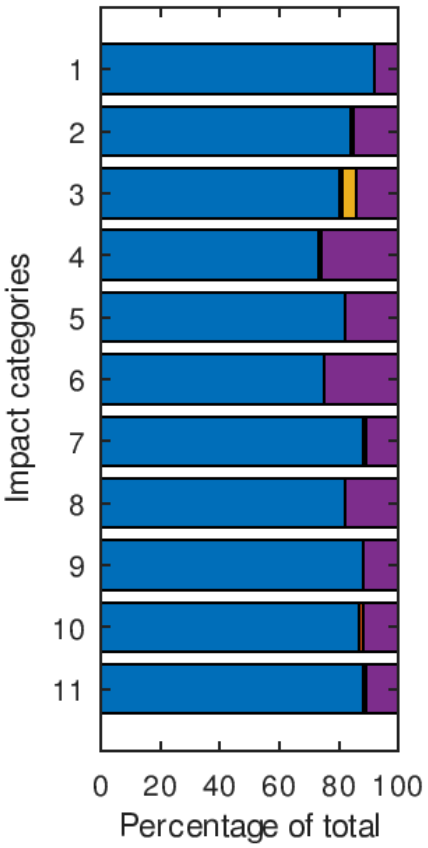
[31] Richardson, K., et al. "Earth beyond six of nine planetary boundaries" *Science Advances* (2023) https://www.science.org/doi/10.1126/sciadv.adh2458

[32] Fang, Kai. "Understanding the Complementarities of Environmental Footprints and Planetary Boundaries." *Environmental footprints* (2021) https://doi.org/10.1007/978-3-030-61018-0_5

[33] Proof: let "a" be the climate change impact energy for fabrications (assumed equal) and "b1, b2" the climate change impact energy per unit of time for use before and after replacing the hardware, "t1, t2, t2' " the durations of use. Replacing the hardware is beneficial for this impact category if a+b1*(t1+t2) > 2a+b1*t1+b2*t2' which simplifies to b1*t2 > a+b2*t2'. Then even with b2=0 we would need b1*t2>a while, as fabrication dominates use, we have a>b1*(t1+t2). Additionally, as t2'<t2 because of the increase in performance, such a replacement encourages extra use leading to a further increase of the climate change impact (rebound effect).

Table 2: Example results from an LCA analysis conducted for the VSIX datacenter at the University of Padova Italy[34]: simplified annual environmental impacts and respective contributions of the fabrication (blue), transportation (red) and usage phases (yellow: coolant leak, purple: energy).

Impact category / Indicator	Value
Abiotic depletion (minerals)	1,54E+01 kg Sb eq
Abiotic depletion (fossil fuels)	1,06E+06 MJ
Global warming (GWP100a)	1,01E+05 kg CO_2 eq
Ozone layer depletion (ODP)	1,03E-02 kg CFC-11 eq
Human toxicity	2,97E+05 kg 1,4-DB eq
Fresh water aquatic ecotoxicity	2,99E+05 kg 1,4-DB eq
Marine aquatic ecotoxicity	4,53E+08 kg 1,4-DB eq
Terrestrial ecotoxicity	4,49E+02 kg 1,4-DB eq
Photochemical oxidation	3,91E+01 kg C_2H_4 eq
Acidification	6,63E+02 kg SO_2 eq
Eutrophication	2,79E+02 kg PO_4 eq



Observing use and analysing inputs from SPECTRUMCoP interviews with providers reveal that users and providers share a growth-oriented vision. This is expected in the scientific and high-performance computing contexts where competition is fierce and much progress has recently been achieved through the increase of resources (larger scientific instruments, scaling of computing infrastructures). However, some interviews mention a desire for harmonising and rationalising uses. The following section identifies gaps and formulates actionable recommendations for developing tools to make this possible. It aims at establishing a framework allowing for managing environmental aspects, both at the local (adaptation) and the global levels (impacts). The core rationale consists in directly using environmental impacts as units of credit given to the users (instead of core-hours) to both constrain their overall footprint and guide them towards responsible use of the resources. This is expected to transfer to the datacenters whose footprint then also becomes tractable so that they can manage, maintain or refurbish their software and hardware along similar principles. Recommendations address three axes for action: reduction of wasteful use of resources, software optimisation and execution tailoring to reduce environmental impacts. As a complement to these technical aspects, policy-orientation recommendations for this management are formulated in the SRIDA, notably as part of Priority 12, to make this possible in a productive way. Both draw from the recommendations, research and development coming from the sustainability-aware computing community.

Beside the environmental impacts, it is worth pointing out that IT also has a number of negative social impacts. Although ISO has published the 14075:2024 norm to quantify those impacts over the whole life cycle, managing these is still very much in its infancy. Although they should not be forgotten, we recommend delaying this. The experience with environmental impacts will prove useful for incorporating the corresponding criteria in a later stage.

[34] Bettiol, M. "La sostenibilità ambientale del digitale: il ruolo dei data center" *Padova University Press* (2023) https://www.research.unipd.it/retrieve/b7595399-7c85-49c6-bcc5-e1ca7bc584c4/9788869383625.pdf#page=64

Gap #1: Current use of IT resources can lead to significant digital waste in the form of poorly used data and software, fragility of execution causing failure after long runs and very low average resource utilisation for scientific pipelines[35].

Recommendation #1: Promote widespread reuse of data and software to reduce duplication of work and allow for cross-fertilisation across projects and scientific communities.

- Support making data and software FAIR to foster reuse at the interdisciplinary level.
- Request justification for the need for new data or analysis.
- Analyse the code semantically to identify functional requirements in order to point to existing libraries and support comparison between alternatives, including estimated environmental impacts for the target architecture.
- Abstract data access through middleware to facilitate use of existing data by new codes and new data by existing software[36]. This middleware could also adjust the data representation based on access patterns to improve the match with the target infrastructure.

Recommendation #2: Improve the resilience of execution by promoting adaptive resource usage by the software to avoid failure due to shortage of resources (power, cooling, compute, storage, duration of reservation).

- Allow for adjusting resource usage of software through adaptive execution strategies for the target platform, notably for parallelisation aspects.
- Develop in-situ analysis to dynamically control execution, including via switching to degraded execution modes.
- Report and aggregate existing and near-future resource usage across on-going software executions to support the datacenter's ability to dynamically adapt its hardware management to changes in resource availability.

Recommendation #3: Develop reference criteria for software technical quality (performance, scalability, energy consumption, reliability, reproducibility) and tooling to measure it automatically to inform developments and evaluate readiness for increased resource usage (e.g. allowing production-level runs).

- Allow quality assessment at selectable granularities as part of continuous integration.
- Estimate how quality is dependent on platform characteristics via comparative evaluations or sensitivity analyses.
- Produce reports and data to track the quality of software throughout its development for developers and infrastructure management.

Gap #2: Scheduling of software execution aims to maximise node occupancy irrespective of environmental impacts.

Recommendation #4: Tailor execution by scheduling jobs based on multiple objectives, including continuum-wide minimisation of environmental impacts.

- Extend the use of "Quality of Service" to fine-tune the job submission process and provide users with means of trading off environmental impacts versus urgency, priority, performance and run and wait times.

[35] The High Performance Conjugate Gradients (HPCG) benchmarks have been designed to better represent the behaviour of scientific applications compared to the High Performance Linpack (HLP) benchmarks. The "Fraction of peak" measurements for HPCG range between 0.3% and 5.5% of the top supercomputers' peak performance (https://www.hpcg-benchmark.org/custom/sc25.html).

[36] Roussel, C. et al. "PDI, an approach to decouple I/O concerns from high-performance simulation codes." (2017) https://hal.science/hal-01587075v1/document

- Based on the requested quality of service, explore the execution parameter space (including energy mix, cooling, efficiency) via mapping to resources with different characteristics and scheduling with time-variable costs[37].
- Provide environmental impact estimates as a result of job submission and execution parameter space exploration prior to execution to allow users to review their submissions.
- Report and carry out accounting based on the environmental impacts of runs.

Recommendation #5: To tailor execution, simulate execution to estimate environmental impacts of different execution scenarios.

- Model the software for off-line analysis of performance and resource usage (for instance in the dataflow paradigm).
- Estimate resource usage and run time corresponding to a mapping and scheduling scenario of the model on a model of the target platform[38] [39].
- Derive environmental impacts estimates from estimated run time and resource usage for the target platform as an input to the mapping and scheduling algorithm.

Recommendation #6: Monitor execution at a low overhead to support estimation of environmental impacts based on precise execution data.

- Record timings and resource usage (including energy and cooling) to estimate environmental impacts live during execution.
- Track environmental impacts to allow the scheduler to manage the jobs and take into account changes in execution conditions (eg. change in the energy mix, issues with cooling and hardware).

Recommendation #7: Improve the prediction of power, energy consumption and corresponding environmental impacts for a given job submission on a wide range of platforms to support where and when to run them.

- Improve modelling of the software-hardware behaviour during execution to support power (instantaneous) and energy consumption (integrated over time) estimation.
- Compare the power and energy consumption across platforms to allow for reliably ranking them based on environmental impacts to support environmental impact-oriented mapping and scheduling.
- Estimate the peak and average power usages to assess compatibility with the infrastructures' power budget.

Gap #3: Tools are missing to explore the design and execution space leading to high environmental impacts linked to poor resource utilisation and wasteful trial and error.

Recommendation #8: Adopt a workflow model for software to assist users to explore the parallelisation design space during development via simulation.

- Support automatically mapping task execution to the available compute (eg. CPU, GPU, other accelerators) and storage resources (e.g. hot or cold storage via abstract data models)
- Support scheduling the different stages in the software to maximise the efficiency of the task distribution and the use of available resources to optimise parallelisation.
- Report inefficiencies for users to improve their software.

---

[37] Carastan-Santos, D. et al. "Scheduling With Lightweight Predictions in Power-Constrained HPC Platforms." *IEEE Transactions on Parallel and Distributed Systems* (2025) https://doi.org/10.1109/TPDS.2025.3586723

[38] Renaud, O., Miomandre, H., Desnos, K., Nezan, J.F.. "Automated level-based clustering of dataflow actors for controlled scheduling complexity." *Journal of Systems Architecture* (2024) https://doi.org/10.1016/j.sysarc.2024.103217

[39] Augonnet, C. Thibault, S., Namyst, R., Wacrenier, P.A. "StarPU: a unified platform for task scheduling on heterogeneous multicore architectures." *Concurrency and Computation: Practice and Experience* (2011) https://dx.doi.org/10.1002/cpe.1631

- Support multiple models for the target architecture and achieve off-line optimisation to generate optimum executables for them based on multiple user-defined objectives, including minimising environmental impacts, as an input to job submission and estimation of environmental impacts.

Recommendation #9: Exploit the workflow model to carry out runtime execution adjustments based on effective behaviour of the software and the availability of resources (including energy).

Recommendation #10: To support optimising the software-hardware match, provide complementary hardware and software views *to developers as a result of test execution.*

- Record resource utilisation and map it to software tasks (eg. via function calls).
- Compute performance and parallelisation metrics and compare them to theoretical limits to guide development.
- Identify power-related hot spots and opportunities for concentrating hardware resource usage with a view to performing impact-performance trade-offs.

4. The Compute and Data Continuum

The compute and data continuum integrates heterogeneous computational resources, federated data systems, and adaptive orchestration frameworks into an interoperable infrastructure optimized for cross-domain scientific workflows. The architecture is designed to enable seamless operation across fast HPC systems, quantum co-processors, and elastic cloud-edge platforms while adhering to FAIR data principles. The following subsections describe the different capabilities, which are defined as usable functions or services that the infrastructure provides to support scientific or data-driven workflows across distributed environments.

The capability map is shown in Figure 1, where the rows denote the principal layers of functionality, from infrastructure resources up to workflow and application services, while columns denote transversal capabilities spanning the entire stack. This distinction is meant to show that some capabilities are part of the system's layered architecture, whereas others provide common support and governance across all layers.

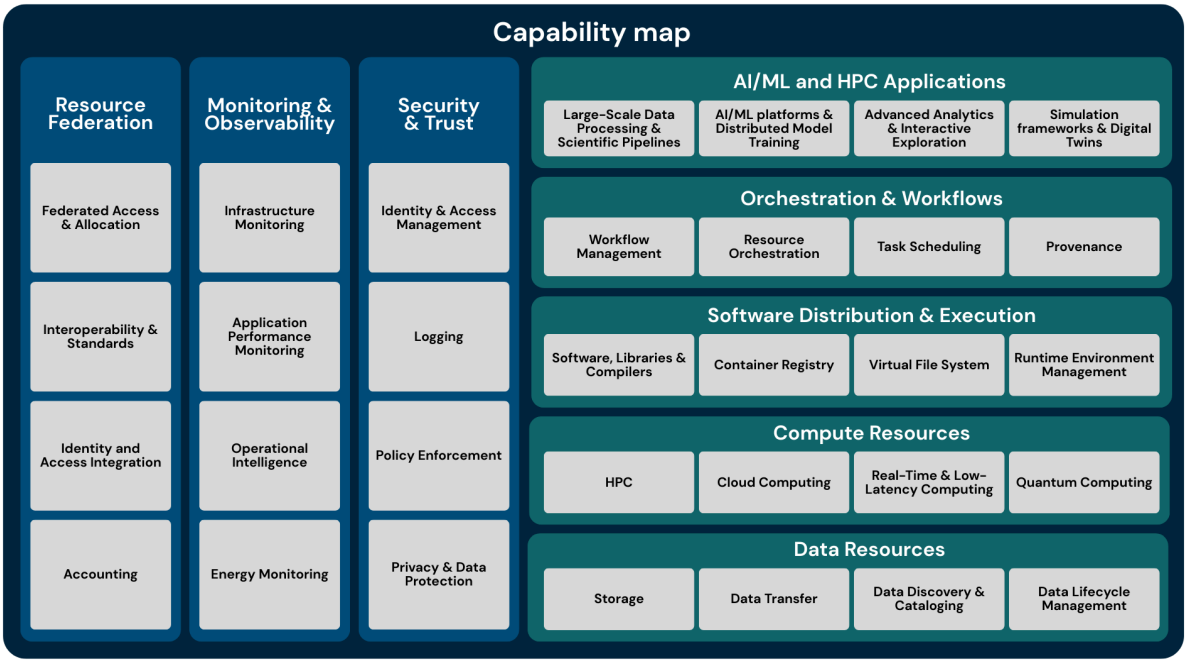


Figure 1: Capability map for the HEP and RA communities.

4.1. Compute Resources

The Compute Resources capability is a foundational layer of the European Compute and Data Continuum, providing seamless access to heterogeneous computational resources distributed across HPC centers, community and commercial clouds, edge infrastructures, and emerging quantum platforms. For data-intensive domains such as HEP and RA, the Compute Continuum transforms a fragmented ecosystem of sites into a federated, performance-optimized environment that can deliver exascale throughput.

HPC and Cloud Computing: As noted previously, HEP experiments have been using HPC resources for more than 15 years, with HPC now making significant contributions to both production and analysis

pipelines[40,41,42,43,44]. To integrate HPC into the continuum, standardized interfaces and APIs must be used so that large-scale batch jobs can be submitted, monitored, and managed across federated HPC centers. This involves bridging local batch schedulers (e.g. Slurm) with global workload managers used by large particle physics experiments, so that simulation and reconstruction tasks can run on HPC without manual effort (as already piloted in HEP). Similar integrations are needed in order to use accelerators such as GPUs transparently, as both CERN and SKAO are exploring heterogeneous computing for AI-accelerated analysis (e.g. GPU enabled queues in SLURM can be made available and accessible via middleware such as DIRAC[45], Panda[46] and glideinWMS[47]).

At the same time, cloud integration offers a complementary pathway for obtaining additional capacity, although the nature of this elasticity depends strongly on the type of cloud, their configuration, policies and utilization. Public hyperscale clouds aim to provide genuine on-demand, elastic resources due to their large overcapacity and commercial operating model, making them suitable for e.g., bursty or exploratory workloads. In contrast, academic and on-premises research clouds, which typically operate at a smaller scale and are often funded to match expected scientific demand, making it less likely that they will always be able to provide immediate, on-demand provisioning. In these environments, cloud interfaces focus more on operational flexibility than guaranteed elasticity: sites may choose to adjust allocations between tenants or VOs, or reserve capacity for specific communities, but overall utilisation patterns often resemble queue-based (HPC-like) scheduling rather than commercial cloud bursting.

Within these constraints, federated cloud stacks (e.g. the EGI Federated Cloud, or EOSC services) still provide important benefits by exposing a unified mechanism to deploy VMs or containerised environments across participating sites. Suitable scientific workflows can then spin up VMs or Kubernetes pods where needed, for example, cloud bursts for Monte Carlo generation or interactive analysis, and shut them down when done (e.g. **CloudComputingElement** from DIRAC). This means that once a scientific workflow is set up appropriately, it can automatically start the computing environments it needs, wherever they are required. Container registries and orchestration services ensure portability across cloud and HPC. Several European efforts (EGI Federated Cloud, EOSC EU Node) already demonstrate cloud virtualization for research. For instance, the DataCloud[48] project showed how big-data pipelines could be deployed across federated cloud sites to stay close to data sources (**egi.eu**). By pooling cloud resources, experiments can handle bursty workloads without long waits for some categories of workloads (e.g. of smaller scale).

Real-Time & Low-Latency Computing: Many HEP and RA instruments produce data at the "edge", e.g. detector electronics or telescope antenna arrays, that must be reduced in size before transfer to central facilities. Edge computing nodes, located near or on-site at observatories and colliders, run real-time pipelines for filtering, compression, or transient detection. For example, LOFAR and the SKAO radio telescope arrays generate raw streams on the order of 10–100 Tb/s[49], which is far too much to send directly offsite, so initial calibration and gridding are done at on-site digital signal processors. By processing data at the source, edge nodes reduce network load and latency. Low-latency analytics (using FPGAs, GPUs or other accelerators) enable rapid responses, for instance immediate follow-up of astrophysical transients. Integrating instrument control systems with edge compute also allows health monitoring of detectors and feedback (e.g. beam diagnostics in accelerators). In summary, edge computing extends the continuum by handling the first stages of massive data flows, ensuring that only valuable information traverses the

[40] Strategy for HPC Integration in WLCG/HEP, Maria Girone et al. (2025) https://zenodo.org/records/15032799

[41] US ATLAS and US CMS HPC and Cloud Blueprint, Fernando Barreiro Megino et al. (2023) https://arxiv.org/abs/2304.07376

[42] Integrating the Perlmutter HPC system in the ALICE Grid, Sergiu Weisz et al. (2025) https://doi.org/10.1051/epjconf/202533701107

[43] Integrating LHCb workflows on HPC resources: status and strategies, Federico Stagni et al. (2020) https://doi.org/10.1051/epjconf/202024509002

[44] Bard, D., Benjamin, D., Calafiura, P., Childers, T., Fisk, I., Girone, M., Gutsche, O., Hufnagel, D., Klimentov, A., Lancon, E., Martinez Outschoorn, V. I., & Wuerthwein, F. (2025). Strategy For HPC Integration (US) Whitepaper. Zenodo. https://doi.org/10.5281/zenodo.17854525

[45] https://doc.grid.surfsara.nl/en/latest/Pages/Practices/gpu_jobs.html

[46] https://npps.bnl.gov/software/panda.html

[47] https://glideinwms.fnal.gov/doc.prd/index.html

[48] https://www.egi.eu/case-study/datacloud/

[49] https://www.skao.int/en/explore/big-data

wide-area network to central archives. This data-reduction at the edge comes at the cost of information loss, but is necessary due to the high data rates. It is simply not feasible to store all raw data.

Quantum Computing: While quantum technologies hold long-term promise, their practical impact on HEP and RA remains limited in the near term. Current exploration is largely experimental, focusing on specific proof-of-concept studies in optimization or simulation. Within the compute continuum, quantum integration is best viewed as a forward-looking interoperability layer rather than a core operational component. While it is desirable to keep the door open for incorporating quantum or hybrid quantum–classical workflows in the future, their practical impact on relevant timescales remains uncertain; the aim is therefore to enable potential integration without disrupting established pipelines.. The emphasis for now remains on readiness and modular integration, rather than active deployment.

Together, the HPC, cloud, and edge components make the computing infrastructure seamless, scalable and flexible, allowing HEP and RA users to run tightly-coupled simulations, large-scale data analyses, and modern AI/ML workloads across the continuum with minimal friction.

4.2. Data Resources

Scientific data in HEP and RA are massive, distributed and federated. The Data Continuum is a federated infrastructure that enables scientific data to be stored, moved, discovered, and preserved across geographically distributed and technologically heterogeneous sites. For data-intensive disciplines such as HEP and RA, where exabyte-scale datasets are generated and shared globally, the Data Continuum is not a peripheral service but a central enabler of scientific discovery. It transforms fragmented storage silos into a seamless, policy-compliant, and performance-optimized ecosystem.

Storage: The continuum relies on a distributed storage capability spanning many sites, which together present a unified namespace to users and applications. Experiments in HEP such as ATLAS and CMS now manage more than an exabyte of data across over 120 storage centers using systems like Rucio[50]. A critical aspect of this capability is data replication and integrity management. Replication rules ensure that each dataset is stored at multiple geographically distinct sites, balancing performance, resilience, and cost. Integrity checks, based on checksums or cryptographic hashes, are performed during transfers and periodically verified in storage to detect corruption or loss. Automated repair mechanisms can re-replicate damaged or missing files from verified copies elsewhere in the federation, maintaining long-term data reliability. Furthermore, access controls and accounting mechanisms are embedded at the storage level to enforce data governance and monitor resource usage across the federation. In practice, HEP and RA communities already share similar approaches: for example, the **ESCAPE project** demonstrated the applicability of Rucio for managing distributed data from the SKAO, CTA, and Vera C. Rubin Observatory, where Rucio is now deployed operationally. Within the continuum, this federated storage capability presents a coherent, policy-driven view of data—users simply request a dataset, and the system locates and retrieves it seamlessly, regardless of where the bytes physically reside. Alternative solutions for Federated Storage are iRODS[51] which is planned to be included in the EuroHPC Federation middleware and Onedata[52] powering the EGI DataHub[53].

Data Transfer: For both HEP and RA, data movement at scale is as critical as compute cycles. Moving multi-petabyte datasets between facilities requires high-performance data movement services and a performant underpinning network. Dedicated transfer nodes and optimized protocols (e.g. XRootD, HTTP TPC[54]) are used to shuttle data reliably. The transfer layer orchestrates transfers, schedules bulk copies to avoid saturating links and queues, and automatically retries failed transfers. For example, current LHC experiments regularly transfer tens of petabytes per day between CERN and remote Tier-1 centers, coordinating thousands of files in flight. Together with dedicated research network backbones such as GÉANT, LHCOPN, and LHCONE, the high-performance fiber infrastructure could ensure that raw and processed data flow efficiently across the continuum.

[50] https://rucio.cern.ch/community/
[51] https://irods.org/
[52] https://onedata.org/
[53] https://datahub.egi
[54] https://twiki.cern.ch/twiki/bin/view/LCG/HttpTpcTechnical

Data Discovery and Cataloging: A critical need for scientists is to find and understand data. The continuum needs to include metadata catalogs and registries to make data (including software) findable. Each dataset should be tagged with rich, standardized metadata (source, date, conditions, format, analysis provenance, processing chain, etc.). Global catalog services keep track of all registered datasets and their physical locations. For example, Rucio provides lookup services and has been used in HEP for years while RA is starting to adopt it. RA also has its own services such as, e.g., the distributed LOFAR long-term archive that provides rich, metadata-based data discovery. More broadly, astronomers make use of the International Virtual Observatory Alliance (IVOA), who have developed standards to make astronomical data and tools interoperable. Provenance tracking is integrated: as data move and transform, metadata records the lineage (which original files were processed by which code version to produce current datasets). This supports the FAIR principles increasingly mandated in research.

Data Lifecycle Management: Scientific datasets have long lifespans and varying usage patterns. The continuum enforces lifecycle policies. For example, raw data might be retained for a fixed period or indefinitely, depending on the need for future reprocessings, intermediate formats deleted sooner, and analysis-ready data preserved indefinitely or refreshed on an annual basis. Automated policies may move data between storage tiers based on access frequency and cost. Quality assurance monitors periodically verify data integrity (checksums, error rates) and rebuild lost fragments if needed. An example in HEP is the ATLAS Data Carousel project[55], which retrieves data from tape to disk regularly to allow for scheduled analyses, and then removes them until next reprocessing. For archival preservation, the system includes curation metadata such as data format, documentation, and software environment, to ensure future usability. In some HEP contexts, these policies are encoded in Rucio rules that replicate datasets N times and expire older replicas when space is needed. RA projects similarly define retention for different data products (e.g. calibrated images vs. raw voltages). The lifecycle block ensures that valuable data remain accessible and trustworthy over decades, while minimizing costs and respecting any legal or safety requirements.

4.3. Software Distribution and Execution

The software, compiler, and library ecosystem is the critical abstraction layer that bridges scientific applications with the heterogeneous resources of the Compute and Data Continuum. While the Compute and Data resources provide the raw infrastructure, it is the software stack that makes these resources usable, and enables portable, and reproducible workflows for global collaborations. For communities such as HEP and RA, whose workflows must run consistently across hundreds of sites and over decades, a robust and harmonized software and execution environment is indispensable.

A central feature of the continuum is the ability to distribute scientific software consistently across federated sites. This is achieved through mechanisms such as software registries, container registries, and virtual file systems.

- **Software and compiler registries** act as catalogues of validated applications, frameworks, and libraries, annotated with metadata on authorship, licensing, and hardware compatibility.
- **Container registries** store portable images (e.g., Docker, Apptainer/Singularity), allowing scientists to encapsulate complex environments and run them reproducibly across HPC, cloud, and edge.
- **Read Only Virtual file systems,** such as CVMFS, enable lightweight, versioned, and up to date distribution of software environments to thousands of nodes, a model already proven in WLCG and elsewhere (e.g. the Vera C. Rubin Observatory)

Together, these services ensure that researchers can access the correct software stack regardless of the underlying hardware or site-specific configuration.

To ensure flexibility, the continuum supports dynamic runtime configuration across heterogeneous systems. This includes environment module systems for selecting and switching software stacks, automated dependency resolution to prevent conflicts, cross-architecture support for x86, ARM, RISC-V, GPU nodes, and other emerging accelerators. Although environment modules are a de facto standard in HPC, they are not yet widely adopted in the HEP/RA communities, which rely more heavily on container-based workflows.

[55] Updates to the ATLAS Data Carousel Project, Mikhail Borodin et al. EPJ Web of Conferences 295, 01054 (2024) https://doi.org/10.1051/epjconf/202429501054

Within the European HPC ecosystem, EasyBuild has emerged as the de facto standard for module generation, with EESSI, distributed via CVMFS and mandated for EuroHPC systems through the EFP, providing an optimised software compilation and distribution layer. Strengthening alignment between HEP/RA software practices and these EuroHPC approaches represents a concrete, actionable step toward improving interoperability and reducing duplication across the continuum.

The execution environment must also support runtime flexibility, enabling users and workflow systems to request specific software stacks, compilers, or libraries declaratively. This includes selecting between different MPI implementations, choosing accelerator-enabled maths libraries, or loading architecture-appropriate builds of AI frameworks such as PyTorch, TensorFlow, or JAX. Exposing these capabilities through stable APIs is essential not only for seamless integration with workflow orchestrators, but also for ensuring long-term reproducibility and reuse. In this perspective, execution environments should be versioned, preservable, and discoverable, so that analyses can be re-executed or adapted over time despite the evolution of hardware and software ecosystems, while avoiding unnecessary redevelopment and duplication of effort.

For both HEP and RA, portable and well-preserved execution environments are key to ensuring that complex reconstruction frameworks, data reduction pipelines, AI/ML models, and simulation codes can run on globally distributed resources with consistent behaviour over long time periods. Technologies such as CVMFS, containerisation, and workflow provenance already play a central role in research software ecosystems and provide a foundation for this continuity[56]. Extending their use to support systematic preservation, reuse, and cross-community sharing of software and its execution context can reduce heterogeneity, improve efficiency, and contribute to sustainability by limiting redundant software development and enabling more effective use of existing computing infrastructures.

4.4. Orchestration and Workflows

The European Compute and Data Continuum will span multiple facilities, resource types, and governance domains. To unlock its potential, scientific workflows must be orchestrated seamlessly across HPC, cloud, edge, and specialized instruments. This requires more than ad-hoc job scheduling; rather, it calls for an integrated orchestration layer that manages resources, tasks, and data flows coherently, while ensuring reproducibility and policy compliance. For both HEP and RA, which rely on large, distributed, multi-decade collaborations, orchestration is what transforms a federation of resources into a functioning continuum.

Workflow Management: Researchers describe analyses as workflows of tasks (data read, calibration, simulation, analysis) with dependencies on data flow and order of execution. A workflow management system captures this DAG of tasks and manages execution end-to-end. These systems also provide resilience, if a job fails at one site, it is retried or moved elsewhere. **Provenance** metadata is recorded automatically for each step, ensuring reproducibility. For example, a HEP experiment could describe an end-to-end simulation chain from event generation through detector simulation to final analysis. The workflow engine then schedules each stage on appropriate resources (HPC, Cloud, CPU, GPU, etc.), and records all details in a database. Software portability could potentially enable the same workflow to be deployed on a HPC center, a community cloud, or an edge device at a telescope without re-engineering. Standards such as CWL (Common Workflow Language[57]) or WDL (Workflow Description Language[58]) provide promising baselines, but must be extended to support federated, heterogeneous execution environments. Within workflows, **task scheduling** ensures tasks are scheduled efficiently, respecting hardware requirements, data locality, and deadlines. For example, a GPU-intensive training job in HEP should preferentially be dispatched to accelerator-enabled nodes, while a transient detection task in RA might have to run on edge resources within milliseconds of data capture.

Resource Orchestration: Beyond describing workflows, orchestration dynamically allocates compute, storage, and network tasks across facilities. This involves mechanisms for dynamic scaling, e.g., expanding into cloud resources when HPC queues are full, and load balancing, sending tasks to underutilized sites to

[56] LOFAR, for example, already deploys its software using CVMFS.
[57] https://www.commonwl.org/
[58] https://openwdl.org/

maximize throughput. Examples are Nextflow, Snakemake, Airflow, Galaxy, etc. For RA in particular, orchestrators must integrate with real-time data streams from telescopes, scheduling compute tasks close to the edge when latency is critical, while offloading heavier processing to HPC or cloud resources.

4.5. AI/ML and HPC Applications

AI/ML and traditional HPC/HTC applications together form the scientific and computational backbone of modern data-intensive research. They drive new methods in simulation, reconstruction, calibration, analysis, and instrument operations, while simultaneously pushing the technological evolution of the continuum. Both classes of workloads benefit from access to large-scale compute, distributed data fabrics, and interoperable software environments, but they expose different performance characteristics that must be supported in parallel. Ensuring that these heterogeneous workloads can operate efficiently across HPC, HTC, cloud, and data-lake infrastructures is therefore a central requirement of the continuum.

Modern scientific workflows increasingly include AI/ML components for pattern recognition, anomaly detection, and surrogate modeling. The orchestration layer must therefore support the full AI/ML lifecycle, including distributed training across HPC/Cloud, federated learning where data cannot leave sites, inference deployment at the edge for low-latency decisions, and model versioning for reproducibility. In both HEP and RA, this could mean deploying pre-trained models at instrument edges to filter events in real time. In HEP, it could also enable large-scale distributed training of generative models or event reconstruction models on exascale resources.

Advanced Analytics and Interactive Exploration: Researchers need tools to explore and analyze large datasets. The continuum offers platforms such as Jupyter notebooks, RStudio, or custom portals linked to the data federation, allowing researchers to query and visualize large datasets through notebooks, dashboards, and statistical tools, bridging traditional batch workflows with real-time exploration. These interactive environments run in the cloud or on HPC visualization nodes, giving scientists graphical and scripting access to data. Statistical and visual analytics libraries (Python, R, machine learning toolkits) are preinstalled. This lowers the barrier for hypothesis generation. Large scale data analysis can start as an interactive process, used to test ideas and solutions on a small-to-medium scale, but can then process as an offline workflow, particularly in cases where the input datasets are PB scale. A seamless transition between the two regimes should be possible, for example via the scale-out of interactive processes from the processing on the user's private machine to large external resources, via tools like Dask.

AI/ML Platforms, Distributed Model Training and Hyperparameter Optimization: AI and ML are becoming indispensable in modern scientific discovery. The continuum provides AI/ML platforms with the tools, infrastructure, and services needed to build, train, deploy, and manage models tailored for large-scale, distributed research. Training infrastructure makes available distributed, GPU-rich training environments spanning HPC, cloud, and edge resources thus enabling models to be trained on massive datasets, including exabyte-scale HEP events or petabyte-scale RA visibility data. Scalable inference services enable deployment of trained models, e.g., at the LHC trigger level to classify collision events, or at telescope edges to detect astrophysical transients within milliseconds. Model registries and versioning play a crucial role. Both AI/ML and HPC workflows benefit from robust software, model, and dataset registries. FAIR-compliant repositories facilitate reproducibility, attribution, and long-term preservation and support collaborative development across geographically distributed communities. For example, an astrophysicist can browse a model zoo for a neural network that classifies radio transients and fine-tune it with local data. Federated learning tools provide privacy-preserving approaches that allow AI models to be collaboratively trained across distributed data without moving sensitive or bandwidth-heavy datasets, an approach particularly relevant for geographically dispersed observatories.

Simulation Frameworks and Digital Twins: Many scientific domains require physics simulations. The continuum hosts domain-specific simulation frameworks (e.g. GEANT4 for particle detectors, N-body or fluid solvers for astrophysics). An emerging class of applications builds on digital twins, which combine physics models, operational data, and AI-assisted prediction to represent instruments or detectors in near real time. Digital twin platforms, such as **itwinai**, allow domain experts to exploit advanced optimisation or multi-node training methods without extensive ML engineering. For RA, digital twins of telescope arrays could for example anticipate calibration drifts or hardware failures while for HEP, digital twins of detectors could for instance model radiation damage or alignment shifts.

Large-Scale Data Processing and Scientific Pipelines: Many HEP and RA applications rely on large-scale, domain-specific processing pipelines that transform raw instrument output into calibrated, analysis-ready data. These include full reconstruction chains, detector calibration workflows, Monte Carlo production, data reduction for RA, and large-scale physics or astrophysics analyses. These pipelines often process petabyte-scale datasets, require sustained compute allocations, and combine CPU- and accelerator-intensive components. Within the continuum, these pipelines must run efficiently on heterogeneous resources and access data through federated mechanisms without requiring per-site customisation. Containerised software environments, standardised data access protocols, and compatibility with heterogeneous architectures are essential to ensuring that these large-scale applications can utilise HPC, HTC, and cloud resources as part of an integrated ecosystem.

4.6. Resource Federation

A European Compute and Data Continuum must allow resources from different providers and domains to function as part of a single, coherent ecosystem. For HEP and RA, where workflows routinely span multiple facilities, this means going beyond isolated access models to enable true **federation of access and allocation** of continuum resources. Researchers should be able to use HPC, cloud, storage, and edge resources through a unified entry point, with allocations that can be honoured seamlessly across institutional and national boundaries.

At the technical level, this requires **interoperable standards** for tasks like job submission, data access, and monitoring, as well as the integration of federated identity and access management systems. Building on existing models such as eduGAIN and the WLCG's Virtual Organisations, the continuum should provide single sign-on across facilities, with support for role-based access, attribute aggregation, and support for service accounts. These mechanisms are essential for ensuring both ease of use and robust security in a federated environment.

Transparent **accounting** is equally important. Usage data, such as CPU/GPU hours, storage consumption, and data transfer volumes, must be collected and reported in a consistent way across all sites. This not only supports fair allocation and equitable participation but also enables funding bodies and infrastructure providers to verify contributions and assess impact. For researchers, transparent accounting builds confidence that resources are distributed fairly, while for policymakers, it provides the accountability needed to sustain investment.

For HEP, this would ensure that exabyte-scale workloads are distributed efficiently across a federation of Tier-0, Tier-1, and Tier-2 centers, with harmonised allocation and usage reporting. For RA, it would allow data from telescopes to flow into European HPC and HTC facilities in a controlled and federated manner, providing astronomers with seamless access to the resources they need for continuous, multi-decade observations.

4.7. Monitoring and Observability

To operate a complex continuum, comprehensive monitoring and analytics are needed. The Monitoring and Observability block ensures that the European Compute and Data Continuum is not only performant and reliable, but also transparent, efficient, and environmentally responsible. For HEP and RA, which operate on multi-decadal timelines and exascale-scale infrastructures, observability and sustainability are not optional features but foundational requirements.

Infrastructure Monitoring: Observability begins with end-to-end monitoring of the health and performance of resources across the continuum. Compute nodes report CPU/GPU utilization, memory and I/O rates while storage arrays report usage and errors. Data center environmental sensors measure temperature, power consumption, and Power Usage Effectiveness (PUE). A centralized telemetry system ingests these metrics in real time. Automated rules detect outliers (e.g. a failed disk, a supercomputer node overheating) so that maintainers can intervene before failure. For example, many HPC sites already use tools like Prometheus and Grafana to visualize cluster health. Over time, trends and predictive analytics will inform

capacity planning (e.g. forecasting that a cluster will be full by next year or that network upgrades are needed). On top of raw infrastructure, **application-level performance monitoring** traces scientific workflows. Each step of a pipeline emits logs and metrics such as job durations, queue wait times, I/O patterns, error counts, etc. Performance monitoring also supports billing and reporting (showing how many core-hours an experiment used). Collectively, these observability data help developers tune code for the continuum and reassure funding agencies that resources are used efficiently.

Operational Intelligence: Raw monitoring data are transformed into actionable intelligence. Analytics and anomaly detection automatically flag unusual behavior, from failing disks to network congestion. Trend analysis and forecasting enable capacity planning, ensuring that infrastructure keeps pace with growing data and compute demands. AI-driven predictive maintenance reduces downtime by identifying hardware likely to fail before it does, a critical feature for 24/7 operations like telescopes or LHC detectors.

Energy Monitoring: Sustainability in the compute and data continuum requires moving beyond simple efficiency indicators toward a more comprehensive, impact-oriented understanding of resource usage. While metrics such as Power Usage Effectiveness (PUE), Carbon Usage Effectiveness (CUE), and water consumption provide useful operational insights, they must be complemented by broader environmental impact assessments, ideally grounded in Life-Cycle Assessment (LCA) methodologies. Fine-grained monitoring of resource usage—covering compute, storage, and data movement—can support more informed decisions, but the objective is not only to optimise efficiency. It is to guide usage toward sufficiency, ensuring that resources are used only where they deliver meaningful scientific value, while avoiding rebound effects associated with unconstrained optimization.

Within this framework, scheduling and execution strategies in the continuum can incorporate environmental awareness as one of several objectives, alongside performance, reliability, and scientific urgency. This includes exploring trade-offs in where and when workloads are executed, taking into account factors such as energy mix, infrastructure characteristics, and overall system impacts. Such capabilities can support more responsible use of distributed resources, improve the robustness and efficiency of data-intensive workflows, and provide transparent accounting of environmental impacts to stakeholders, including the European Commission and member states.

4.8 Security and Trust

The European Compute and Data Continuum will be a federated infrastructure spanning multiple countries, institutions, and administrative domains. In such an environment, trust and security are not add-ons but foundational enablers (e.g. **AARC Blueprint Architecture**). They ensure that (sensitive) data, compute resources, and workflows are protected while still enabling frictionless collaboration across organizational and national boundaries. Both HEP and RA rely heavily on global collaboration, but they have to comply with EU regulations on privacy, data protection, and cybersecurity.

Identity and Access Management (IAM): Access to resources is governed by a federated IAM system. HEP and RA communities rely on global identity federations (such as eduGAIN) so that researchers can use their home institution credentials across all sites. Once authenticated, user attributes (collaboration membership, VO[59] roles, project grants) travel with the identity so that access policies can make decisions. For example, an authenticated physicist's request for a dataset triggers a check in the federated IAM: if he/she is a member of the owning experiment, access is granted. According to the survey carried out by the SPECTRUM CoP there is a lot of fragmentation in the currently used tools, almost equally shared between Indigo-IAM[60], CERN-SSO[61] and EGI Check-in[62] (with MyAccessID being used by the EFP). A unified solution would be preferable for any future compute and data continuum.

Policy Enforcement: Trust in a federated environment depends on shared rules, policies, and validation processes. The trust framework must define the criteria for participation (technical, organizational, and policy compliance) and provide mechanisms for continuous validation of participants. For example, sites

[59] Virtual Organisations.
[60] https://github.com/indigo-iam/iam
[61] https://auth.docs.cern.ch/
[62] https://www.egi.eu/service/check-in-internal/

must demonstrate that their infrastructure complies with minimum security baselines, privacy regulations, and FAIR principles. Digital certificates and accredited attribute providers reduce reliance on manual validation, making trust decisions scalable (however many modern legal interpretations require that all actions must be traceable to a unique and natural person and this needs to be solved for bulk workloads). In practice, this mirrors existing trust models in the WLCG but extends them to include HPC and cloud facilities with different compliance regimes. Once policies are defined, they must be enforceable in real time. Policies define rules such as "data from telescope X can only be stored in GDPR-compliant facilities within the EU" or "users must re-authenticate with multi-factor authentication when accessing Tier-0 resources."

Privacy and Data Protection: While most HEP and RA data are not personal, the continuum will inevitably process some sensitive datasets (e.g., medical imaging in cross-domain digital twin applications, AI models trained on personal mobility data for environmental research). Compliance with GDPR and future EU data protection legislation is mandatory. Technical safeguards include encryption at rest and in transit, differential privacy techniques for sensitive AI/ML use cases, federated training of AI models, and logged trails for data access. The framework must also respect data sovereignty, ensuring that data remains within designated jurisdictions when required.

5. Technical activities and recommended actions

The development of a European Compute and Data Continuum presents a series of interdependent technical challenges. A key requirement is the establishment of standardized interfaces that facilitate interoperability between diverse systems and enable efficient use of advanced architectures, such as GPUs, without necessitating extensive custom integration at each computing center. Equally important is the development of robust data pipelines capable of managing massive throughput while adapting to the dynamic resource availability typical of HPC environments. Evolving workflows, particularly those incorporating machine learning, offer opportunities for accelerated scientific discovery but also necessitate significant efforts in code reengineering, performance optimization, and advanced scheduling. Additionally, strict security and authorization requirements at HPC facilities must be addressed in a manner that maintains data integrity and aligns with the existing HEP trust models .

Similarly, the next generation large RA projects (LOFAR2.0 and the SKAO) are facing the challenges associated with processing significant quantities of data across heterogeneous architectures. In the case of the SKAO, the final data processing steps and the interface with the international scientific community will be in the form of distributed nodes of an SRCNet. Currently smaller in scale than the WLCG, the operations model of SRCNet is similar in that countries, or institutes contribute in-kind resources to run their local nodes. Many of the same requirements and hurdles to be overcome by WLCG and HEP are common to LOFAR2.0 and SRCNet. The following sections propose strategies for achieving the development of a high-performance, efficient, and secure European Compute and Data Continuum.

5.1. Standardization of Interfaces

Related Capabilities	Resource Federation, Monitoring & Observability, Security & Trust, Orchestration & Workflows, Software Distribution & Execution, Compute Resources, Data Resources, AI/ML and HPC Applications
Related Requirements	2, 4, 5, 6, 15

As established in Section 2, today, each supercomputing center uses its own portals, schedulers and software stacks. Therefore, an experiment wishing to run on multiple HPC centers must custom-adapt to each.

HPC centers operate with distinct access protocols, policies, and resource management systems, reflecting their diverse operational requirements. While this heterogeneity serves the needs of individual centers and their specific user bases, it poses significant obstacles to integrating these resources into distributed infrastructures such as WLCG and SRCNet. Researchers must manually adapt their workflows to each site's unique requirements, a process that is inefficient, error-prone, and, in some cases, infeasible. Without a standardized approach, scaling the number of participating HPC sites remains a labor-intensive challenge. A concerted effort involving funding agencies, WLCG, SRCNet, HPC organizations, and resource providers is required to define and implement a minimum set of standard interfaces. Collaboration with initiatives such as the EuroHPC JU provide promising avenues for achieving this goal.
The HEP and RA communities should adopt or develop middleware that presents interoperable interfaces. In addition, collaboration with EuroHPC's EFP to adapt its interfaces would simplify the process: the EFP is defining a federated identity and allocation API (with single-sign-on and certificate flows). Aligning with such efforts means WLCG and SRCNet users could obtain access tokens usable across all participating systems. For RA, the SRCNet can benefit strongly from the integration with these standards so that SKAO pipelines can run on any compliant HPC.

The increasing diversity of user groups, including those unfamiliar with classical supercomputing or accustomed to different tools and paradigms, with artificial intelligence being a prime example, highlights the need for more diverse and more accessible interfaces.

The EFP architecture emphasizes modularity, flexibility, and API-centricity as core design principles while being minimally intrusive to existing systems. The platform consists primarily of open source components that are integrated to provide necessary connections to federated systems. Examples of key components which are aligned with HEP/RA needs include EFP Authentication and Authorization Infrastructure (AAI) leveraging MyAccessID, EFP Software Catalogue implementing EasyBuild and EESSI via CVMFS for consistent software distribution across systems.

In summary, computing centers currently expose a variety of site-specific interfaces (portals, schedulers, credentials) that complicate integration with distributed infrastructures (WLCG, SRCNet) and require experiments to adapt their workflows to each site. This fragmentation is inefficient and limits scalability. Key findings include the need for a common middleware or API layer and federated authentication/allocation mechanisms.

- **Action:** Develop or adopt standardized interfaces for HPC access, common across all European compute sites. Uniform interfaces would benefit HEP and RA communities and HPC centers through reduced integration effort.
- **Action:** Integrate HPC centers with middleware interoperable with EuroHPC Federation AAI and allocation APIs. This way users could authenticate once via their institutional or VO credentials and their project allocations would be automatically recognized and enforced across sites.

5.2. Co-Design of Computing, Data, and Networking Infrastructure

Related Capabilities	Compute Resources, Data Resources
Related Requirements	6, 7, 12, 13

The successful deployment of compute and data continuum capabilities for HEP and RA requires careful co-design of computing, data management, and networking infrastructure to ensure optimal performance and cost-effectiveness. This co-design approach must consider the specific characteristics of scientific workloads, data access patterns, and collaboration requirements that distinguish these domains from commercial and other scientific computing applications.

HPC systems are typically procured based on performance metrics and benchmarks and the needs of established computational disciplines, including lattice quantum chromodynamics (QCD), computational chemistry, astrophysics, and climate modeling. However, the architectural choices often made to support these workloads, such as limited RAM, minimal or absent local storage, restricted IP networking, and highly controlled external connectivity, can render certain HPC systems impractical for HEP workflows. In some cases, technical workarounds can enable partial integration, but these solutions are site-specific, require substantial R&D, decrease workflow efficiency, and introduce significant maintenance burdens. To ensure long-term compatibility, HEP and RA must engage with funding agencies during the design and procurement phases of HPC systems. For instance in the case of EuroHPC, the most effective way to be part of the co-design process would be to obtain a recognized status within EuroHPC and in the long term seek strategic access to EuroHPC resources. On top of guaranteeing a stable share of the resources, it would enable the domains to be part of the design and requirement definition for new EuroHPC systems.

Network infrastructure represents a critical constraint for exascale scientific computing, as the ability to move data efficiently between storage and computing resources often determines overall system performance. The integration of dedicated research networks, commercial cloud connectivity, and specialized links between major facilities requires careful optimization to support diverse traffic patterns while maintaining security and cost-effectiveness. On this topic, GÉANT has been awarded a contract by the EuroHPC JU to deliver high-capacity, secure, pan-European connectivity for Europe's supercomputing infrastructure. This hyperconnected network aims to connect selected HPC centres, national supercomputers, AI factories, quantum facilities, and research and data centres across the continent, enabling seamless collaboration for researchers, industry, and the public sector. The project covers the design, implementation, and operation of the hyperconnectivity infrastructure over a period of 48 months

In summary, most current European HPC systems are optimized for other scientific domains and often have limited memory, ephemeral storage, and restricted I/O, features that hinder HEP/RA workloads. Key findings emphasize that HEP/RA communities must influence system design, become recognized stakeholders in national procurement (and ideally EuroHPC), and embed their requirements early. For instance, the Destination Earth climate initiative secured multi-year exascale access via strategic EuroHPC support demonstrating how strategic partnerships can secure needed resources.

- **Action:** Seek official "strategic access" or long-term allocations in EuroHPC/national HPC programs. This would allow HEP and SKAO communities to gain stable compute shares and a voice in new system design.
- **Action:** Fund initiatives on software-defined, programmable networking to enable intelligent, high-performance data movement.

5.3. Software Portability and Heterogeneous Architectures

Related Capabilities	Software Distribution & Execution, Orchestration & Workflows, AI/ML and HPC Applications, Compute Resources
Related Requirements	4, 5, 11

Both HEP and RA rely on extensive and highly specialized codebases that have often been developed over many years by large collaborations. Historically, these applications have been optimized for x86_64 CPU architectures, which dominated HEP and RA computing resources. In contrast, modern HPC centers increasingly emphasize heterogeneous computing architectures, particularly GPUs, for their superior energy efficiency and computational throughput. Efficient utilization of these accelerators requires substantial modifications to existing software, including parallelization strategies tailored to specific hardware. While portable programming models and libraries have made progress in easing this transition, significant challenges remain. Existing AI/ML workflows are often the first to adopt GPUs due to their naturally parallel nature and reliance on well-supported libraries like TensorFlow or PyTorch. This is true for RA workflows required to create advanced data products such as astronomical source lists, where AI models have become increasingly common. However, general-purpose event simulation, reconstruction, and analysis codes remain largely CPU-bound. Continued reliance on x86_64 CPUs restricts HEP applications to architectures that are no longer advancing as rapidly in terms of performance and energy efficiency. Without addressing portability, HEP risks being unable to leverage cutting-edge HPC resources, limiting scientific output and inflating computing costs. HEP experiments are currently piloting several development efforts to offload simulation and reconstruction codes to accelerators, but in general no breakthrough/significant speed ups were observed. Rather, these serve as being able to run the codes on a wider variety of hardware. A large fraction of CPU will continue to be required to control the accelerators pipelines.

Modern and legacy applications require access to configuration files and software environments, often distributed through CVMFS for HEP. A comprehensive survey among HEP experiments is needed to assess their readiness for accelerated architectures, evaluate the feasibility of adopting portability frameworks, and identify high-impact pilot projects for heterogeneous computing. Establishing a mechanism for continuously monitoring the evolution of GPU usage and its impact on computing models will be essential for future-proofing HEP software infrastructure.

In summary, HEP/RA software stacks have been developed for HTC workflows using CPU-only platforms, but modern HPC emphasizes accelerators (GPUs, AI chips). A key finding is that, without adaptation, HEP/RA will under-utilize future resources. Portable programming models and containerization are critical to decouple code from hardware. Experiments must identify which applications (simulation, reconstruction, etc.) can exploit accelerators and refactor them accordingly.

- **Action:** Fund surveys of experiment codebases for accelerator and HPC readiness and identify priority applications for adaptation.
- **Action:** Fund pilot projects to port critical codes using portability frameworks and optimize HPC exploitation. Produce optimized prototypes and best practices.

- **Action:** Use CVMFS and/or containers to encapsulate complex software for distribution. This includes GPU-enabled ML frameworks (TensorFlow, PyTorch) and related libraries to ensure AI-driven workflows run efficiently on both grid and supercomputers.
- **Action:** Support curation of software and associated documentation (metadata , repos etc) following open science policies and FAIR best practices.

5.4. Data Management and Network Performance

Related Capabilities	Data Resources, AI/ML and HPC Applications
Related Requirements	6, 7, 8, 9, 12

Efficient data access is critical for HEP and RA applications running on HPC systems, as data handling directly affects the ability to fully utilize computational resources. While experiments have developed advanced data management systems, HPC facilities were not designed for data-intensive workloads, creating challenges that must be addressed to ensure performance at scale.

Another major constraint is network connectivity. Data must be moved in and out of HPC facilities at a rate that matches compute speed, which makes reliable, high-throughput networking essential. At minimum, HPC sites need strong connectivity to WLCG centers and SRCNet nodes, which needs to be realized via e.g., the GÉANT Network. However, there are currently no common tools for automating large-scale data transfers between WLCG, SRCNet, and HPC sites , forcing reliance on solutions offered by individual HPC centers which differ from site to site and which are not optimised for the data transport needs of HEP/RA. This complicates integration and raises concerns about scalability.

The current limitations prevent major uptake of HPC sites in HEP and RA, and only a portion of the available HPC sites in Europe are currently exploited. Future growth in data volumes and processing demands makes it critical to develop standard, open interfaces for data transfer and storage access at HPC sites so as to fully exploit them. Close attention must also be paid to site policies, such as restricted general connectivity from compute nodes, which could further constrain operations.

In summary, HEP/RA workflows handle multi-petabyte datasets, but HPC sites rarely provide long-term archival storage. Their file systems are optimized for short-term high-speed use (in most cases as high performance scratch areas), so large data volumes must be staged in and out for each campaign. Key findings stress that scalable data integration requires robust, high-bandwidth networks and automated transfer tools.

- **Action:** Conduct combined SRCNet-WLCG-LOFAR-HPC data challenges to validate end-to-end workflows and identify bottlenecks under realistic loads.
- **Action:** Establish a coordinated programme to develop and deploy standard, high-throughput data transfer tools and network integration across WLCG, SRCNet, LOFAR, and HPC sites, interoperable with the EFP. This should include the design of common data movement interfaces, performance tuning over e.g., the GÉANT backbone, and shared deployment of optimized transfer services (e.g. FTS, Rucio, iRODS) to replace current site-specific solutions.
- **Action:** Fund projects to harmonize policies across computing centers to facilitate integration with HEP/RA and other communities' workflow orchestration, and prove interoperability via data challenges.

5.5. Workflow Adaptation and Optimization

Related Capabilities	Orchestration & Workflows, Software Distribution & Execution, Resource Federation, Security & Trust, Monitoring & Observability
Related Requirements	1, 2, 3, 5, 15

HEP and RA experiments rely on complex, diverse workflows to manage tasks such as data processing, reprocessing, Monte Carlo event generation, detector simulation, and physics analysis. RA workflows are similarly complex. In the case of HEP for instance, these workflows, developed over decades and representing significant community investment, are fundamental to interpreting experimental data, simulating particle interactions, and maximizing physics output. Well-established workflow systems handle the majority of these computational needs, but integration with HPC centers remains limited.

To improve HPC usability, a more uniform mechanism for integrating experiment workflows with HPC resource management systems is necessary. Given the diversity of tasks, workflows must be adapted to exploit the computational models of HPC environments while preserving their scientific objectives. Alternatively the Computing Elements solutions currently exploited by HEP/RA (ARC-CE/HTCondor-CE) could be configured to access HPC systems (this is already done for instance by the ATLAS experiment with EuroHPC Vega[63]).

Time-Sensitive Patterns: Time-sensitive workflows require rapid processing and responsiveness, driven by needs such as fast event scouting, rapid re-training and tuning of real-time machine learning models, and data quality monitoring. These tasks depend on immediate data access and low-latency compute cycles, making their execution on HPC centers challenging under current conditions.

Data-Integration-Intensive Patterns: Another class of workflows is characterized by large-scale data integration and processing. These include detector calibration, data reduction, and complex analyses that extract physics results from extensive datasets, often geographically distributed on the global scale. Efficient execution of these workflows depends not only on computing power but also on robust data handling and storage capabilities, which are often constrained by HPC policies and architectures.

Long-term Campaign Patterns: Finally, long-term campaign workflows dominate HEP computational demand. These campaigns span years and include event generation, detailed detector simulations, repeated event reconstruction, and derivation of analysis-ready datasets. End-to-end processing, meaning sequentially executing the full chain of workflows, from event generation to final data derivation or analysis, is very important to minimise data flows to and from the HPC. Similarly for RA, data processing for large key science programmes generally involves the rerunning of multiple workflows as new observatory data products become available. Such tasks require substantial, sustained computing resources and benefit most from platforms that support long-term engagement and continuity. HPC centers that lack policies supporting persistent, multi-year campaigns are a poor match for these workflows.

At present, only a few HEP workflows have been optimized for HPC architectures, data management models, and authorization systems, typically tailored for a single HPC center. Policy constraints, such as the lack of general internet connectivity from compute (worker) nodes, further complicate integration. This limits the ability to run pilot-based workload management systems that rely on dynamic, late scheduling of tasks to resources.

Fully leveraging HPC centers requires the ability to execute a broader range of HEP workflows, particularly those with the highest computational demands. Without this capability, workflow selection becomes more complex and resource utilization suffers. Moving forward, collaboration between experiments and HPC initiatives is essential to identify workflows best suited for HPC execution under current constraints. This includes assessing the impact of technical and policy limitations and developing strategies to mitigate them. Establishing such a consensus will be critical for maximizing the scientific benefit of HPC resources for HEP.

[63] https://link.springer.com/article/10.1140/epjc/s10052-024-13701-w

In summary, HEP and RA workflows are complex, multi-stage systems developed over decades to support data simulation, reconstruction, and analysis. However, their integration with HPC centers remains limited due to differing software environments, network policies, and scheduling models. To exploit HPC effectively, workflows must be adapted to these environments and coordinated across experiments and HPC providers.

- **Action:** Integrate existing application specific workflow systems with HPC resource managers to enable native submission, data staging, and monitoring across federated infrastructures to allow for orchestration of complex workflows across grids and supercomputers.
- **Action:** Co-design and converge on a set of common APIs for workflow systems, with support for unattended execution, provenance tracking, and cross-infrastructure orchestration.

5.6. Security, Authorization, and Authentication

Related Capabilities	Security & Trust, Resource Federation
Related Requirements	2, 3, 5, 9

HEP collaborations rely on the WLCG Trust model, which enables sites to grant access to users and workflows based on their membership in federated Virtual Organizations (VOs). Similarly, SRCNet is also adopting AARC-BPS FAAI. This approach simplifies user management but presents challenges when integrating with HPC centers that impose stricter authentication and authorization policies. Unlike WLCG, most HPC centers do not support VO-based access control, nor do they have federated identity management systems capable of handling dynamic authorization via entitlements. Current practices require individual users to obtain site-specific credentials, limiting workflow automation and increasing administrative overhead. In addition to this, HPC centers privilege, if not require, accounts to be linked to identifiable individuals, and do not encourage the use of service-related accounts.

While tolerable for limited workflows, the absence of a federated model ultimately constrains the efficient use of HPC resources by large collaborations. Addressing this gap requires the development of a security framework that balances the global, distributed nature of HEP with the specific security requirements of HPC centers.

The foundation of such a solution would be a federated Authentication and Authorization Infrastructure (AAI) based on industry standards. This would enable users to authenticate once and access multiple sites securely, with consistent authorization policies. Building such an infrastructure is essential to facilitate international collaboration, ensure compliance with regulations, and fully integrate HPC resources into the HEP computing ecosystem.

The EFP is starting to fill this gap by enabling MyAccessId and Edugain Access to EuroHPC systems, therefore the current AAI system used by HEP/RA (Indigo-IAM, an AARC-BPA compliant system supporting OIDC protocol) will need to be made interoperable. A notable missing piece in the EFP is the support for service accounts, which is often required by the HEP/RA communities to enable long running services and central processing workflows.

In summary, HEP and RA collaborations rely on a VO-based trust model and X.509 certificates, whereas HPC centers typically require independent accounts or federated identities. Key findings highlight the need for a unified AAI framework: users should authenticate once and have their roles recognized everywhere. The ability to verify the certificates of new nodes being added to a federated compute network is also critical.

- **Action:** Build a federated AAI that maps VO membership to HPC access control to enable users to sign in once and gain appropriate privileges on all sites, interoperable with the EFP. Multiple AAI systems can coexist, if they are AARC-BPA compliant.
- **Action:** invest in translation services which map identities from different AAI infrastructures, when no direct interfacing is available..
- **Action:** Harmonize HPC center authorization policies to accept federated identities to eliminate the requirement of per-site account management.

- **Action:** Develop standardized support for service accounts within the federated AAI framework. HEP/RA communities should work with the EFP, HPC sites, and AAI providers to define a secure, auditable model for service accounts that aligns with both HEP/RA automation needs and HPC security policies.

5.7. AI/ML Integration and Computational Trends

Related Capabilities	AI/ML and HPC Applications, Compute Resources , Data Resources, Software Distribution & Execution
Related Requirements	4, 5, 8, 11, 12, 13

AI and Machine Learning (ML) techniques present significant opportunities for HEP and RA, enabling new approaches to simulation, event selection, reconstruction, and data analysis. These methods can outperform traditional algorithms in both efficiency and scalability, accelerating scientific progress.

The growing role of GPUs in ML applications is particularly relevant, driven by substantial industry investment in AI-optimized hardware. However, this shift presents challenges for HEP. The industry focus on accelerating low-precision computations for AI comes at the expense of double-precision performance, which remains critical for many traditional HEP and RA workflows, but also for other domains. As this trend continues, each experiment will need to assess which computational tasks can be refactored to leverage AI/ML techniques and which will require specialized hardware capable of high-precision calculations.

These challenges highlight the importance of close collaboration between data-intensive scientific communities and HPC system designers. Engaging at the design stage is essential to ensure future computing resources can adequately support both AI-driven workflows and the precision requirements of classical HEP computations.

In summary, AI and Machine Learning are becoming increasingly important tools in HEP/RA (for simulation, reconstruction, analysis). In parallel, HPC hardware is shifting toward AI-optimized, low-precision accelerators. Key findings note that experiments must strategically refactor suitable tasks for ML while planning for high-precision workloads. Close collaboration is needed so that future HPC procurements include both powerful AI accelerators and sufficient double-precision compute.

- **Action:** Identify analysis and simulation tasks suitable for ML acceleration and fund and coordinate development of AI-driven workflows.
- **Action:** Communicate precision and performance needs to HPC hardware planners to ensure new systems have a balanced hardware mix. This applies not only to HEP/RA but other scientific domains as well (digital twins, traditional simulation).
- **Action:** Train physicists and astronomers in ML theory, frameworks and tools to build the capacity to fully exploit AI factories. Cross-domain training is encouraged.

5.8. Long-Term Resource Provisioning and Accounting

Related Capabilities	Monitoring & Observability, Resource Federation
Related Requirements	1, 10, 13

HPC resource allocations are typically awarded through peer-reviewed proposals with usage limited to a fixed period, often one year or less. This model contrasts with the long-term nature of experimental activities, where complex instruments are operated over multiple decades, with sustained computing demands. The LHC computing model, for example, relies on multi-year resource commitments formalized through Memoranda of Understanding (MoUs) with WLCG sites, ensuring continuous availability and predictable resource planning aligned with physics goals.

While experiments invest in portable software capable of running across diverse platforms, the lack of long-term allocations at HPC centers limits the return on further adaptation efforts. Multi-year or renewable annual guaranteed allocations would enable experiments to fully leverage such investments and integrate HPC resources more effectively into their computing models.

Strategic science initiatives, such as DestinE, recognized by the European Commission, benefit from stable access to HPC resources without yearly reapplication. HEP and RA aim to strengthen their engagement with EuroHPC and national HPC programs by responding systematically to access opportunities, including development, benchmarking, regular and extreme-scale access, AI, and data-intensive application calls, while building the internal capacity to position itself as a candidate for future strategic access.

In summary, LHC and SKAO observing programmes can span many years, but HPC allocations are typically granted for one year. Key findings stress that multi-year or easily-renewable allocations are essential. HEP/RA should aim to secure long-term commitments and ensure usage is tracked alongside WLCG/SRCNet resources.

- **Action:** Develop multi-year and strategic HPC projects to secure guaranteed compute allocations aimed to create stable capacity aligned with experimental timelines.
- **Action:** Fund a study to propose practical models for multi-year or renewable HPC allocations aligned with experimental lifecycles.

6. Conclusions

The SPECTRUM Technical Blueprint demonstrates that Europe's scientific ambitions in HEP and RA would greatly benefit from a fully integrated compute and data continuum. Such an ecosystem would bring together HPC, HTC, cloud, edge, and emerging technologies (including eventually quantum systems), but its success depends on addressing a series of fundamental challenges. This Technical Blueprint has outlined a comprehensive vision for the European Compute and Data Continuum, designed to meet the needs of data-intensive scientific domains such as HEP and RA. It translates community evidence into a modular design that structures the continuum around key capabilities, defining the technical foundations required for Europe to maintain leadership in scientific discovery. Even if targeted to the HEP and RA communities, the findings and recommendations are common to other computing- and data- intensive scientific workflows, especially if geographically distributed; the blueprint, hence, must be considered in more ample terms than those limited to the specific domains.

A number of consistent themes emerged during the creation of the blueprint and the previous work done in SPECTRUM. A first finding is the need for standardization of interfaces and that federating capabilities and interoperability are indispensable. Today's infrastructures remain fragmented, with each site exposing distinct APIs, protocols, and operational practices. This hinders interoperability, complicates cross-border workflows, and makes reproducibility fragile. Establishing open, community-driven standards for job submission, data access, monitoring, and accounting is therefore essential. Only through common interfaces can the continuum become more than the sum of its parts.

Second, is the co-design of compute and data infrastructures. HEP and RA workflows do not separate computation from data movement, both are tightly coupled, often requiring massive parallel compute and transfers in the same workflow. The continuum must therefore be co-designed with balanced investment in compute, storage, and networking, ensuring that bottlenecks are removed and resources evolve in harmony with scientific requirements. Engagement between scientific collaborations, HPC centers, and network providers must become a structured, long-term process. Although much work has already been done in the HEP community to integrate HPC resources with experiment workflows both in Europe and in the US more effort is required to address the growing needs of RA.

The third challenge concerns software portability across heterogeneous architectures. As Europe invests in diverse hardware such as GPUs, FPGAs, ARM processors and quantum accelerators, scientific codes must be able to run efficiently across all or most of them. Current software stacks are often bound to specific architectures, limiting flexibility and increasing maintenance costs. Investing in portable libraries, containerized environments, and compiler toolchains that target multiple backends is thus a priority. For communities like HEP and RA, where workflows must remain operational for decades and hence generations of computing systems, portability is of utmost importance.

A fourth key finding is the importance of workflow management. Scientific research increasingly depends on complex, multi-step workflows that span facilities and scientific domains. These must be orchestrated seamlessly across the continuum, with robust scheduling, provenance tracking, and error handling. Without such workflow intelligence, the continuum risks remaining fragmented and underutilized. It is critical that these workflows be run while following FAIR data practices (including metadata and provenance capture for reproducibility).

Security and trust also emerge as essential. Authorization and authentication mechanisms remain uneven across infrastructures, creating barriers to collaboration. A federated approach to identity and access management, building on standard initiatives supported by the community, is needed to ensure that researchers can move seamlessly between resources. Security must be embedded at the architectural level, not retrofitted as an afterthought.

The Technical Blueprint highlights the transformative role of AI/ML integration. Machine learning is no longer peripheral to scientific discovery, it is embedded in many parts of the HEP and RA communities production and analysis workflows. The continuum must therefore provide AI/ML services at scale, including distributed training environments, inference platforms, and model registries. Supporting and providing these capabilities ensures that AI accelerates, rather than complicates, scientific workflows.

Finally, long-term sustainability is a defining concern. Resource provisioning and accounting must move from short-term, project-specific allocations to stable, predictable, multi-year commitments. Transparent accounting mechanisms should span HPC, HTC, cloud, and data infrastructures, enabling tracking usage as well as energy and carbon costs. Ensuring a skilled and adaptable workforce through continuous training and knowledge transfer is equally essential to maintain European technological and scientific autonomy over time. For HEP and RA, whose scientific lifecycles extend over decades, this long-term perspective is the only way to guarantee reproducibility, scientific continuity, and strategic autonomy.

Taken together, these findings define both the challenges and the opportunities. Europe has the chance to build a continuum that is open, federated, and sustainable and adherent to its standards of equity, inclusion and attention to citizens' rights. This will require technical innovation in standardization, portability, workflows, and AI integration, but also organizational innovation in governance, funding, and long-term provisioning. If Europe acts now, researchers will soon be able to, for example, transparently combine GPU nodes in France, cloud services in Italy, and edge devices in Spain within a single workflow. Such a future will not only secure the scientific leadership of Europe within HEP and RA, but also establish a model digital infrastructure for all data-driven disciplines.

Next steps should focus on piloting and validation. Technically, prototypes and pilots must validate the blueprint outlined here, ensuring interoperability across domains and readiness for production use. Organizationally, governance and funding models must be agreed upon at European and national levels, embedding long-term commitments that guarantee continuity. Without these steps, the blueprint will not come into operational reality.

In conclusion, the Compute and Data Continuum is not simply an infrastructure plan but a strategic investment in Europe's scientific future. By integrating technical excellence with policy alignment and sustainability, it provides the foundation for discoveries at the energy and cosmic frontiers, and a model for data-driven science across disciplines.

Appendices

A Related Initiatives

JENA **(Joint ECFA–NuPECC–APPEC) Activities** brings together the particle, nuclear, and astroparticle physics communities to coordinate cross-domain challenges. Its working groups have highlighted shared needs in areas such as simulation frameworks, detector software, and data analysis platforms. JENA has been particularly active in promoting sustainable software practices, including the adoption of common development standards and training curricula for research software engineers. The initiative also stresses the importance of long-term funding for software and infrastructure, recognising that major scientific programmes can span decades. JENA's focus on human resources, through career pathways and skills development, echoes needs expressed by the SPECTRUM Community of Practice.

InPEx **(International Post-Exascale Project)** is an international collaboration among HPC researchers from Europe, the United States, and Japan. It serves as a forum for discussing challenges and opportunities in the exascale and post-exascale era. Topics include hardware–software co-design, HPC–AI-Quantum convergence, the creation of open machine learning models and datasets, FAIR data stewardship, energy efficiency, and ethical considerations around AI. InPex produces reference documents that inform investment strategies and encourage harmonisation of HPC development across regions. Its global perspective contributes to Europe's approaches remaining interoperable with international infrastructures and aligned with worldwide best practices, which is critical for globally collaborative sciences such as HEP and RA.

Strategy for Post Exascale **(SPE)** is an EU funded project that focuses on preparing the convergence of High PerformanceComputing (HPC), Artificial Intelligence (AI), and Quantum Computing and Simulation (QCS). It will identify key research and technological challenges and produce a strategic roadmap, as well as policy papers to guide Europe's research and innovation agenda beyond exascale. The initiative fosters cross-community collaboration in Europe and beyond through strategic exchanges with similar worldwide initiatives contributing to the development of a European competitive HPC/Quantum/AI ecosystem.

ETP4HPC's Transcontinuum Initiative **(TCI)** is a European effort to articulate and advance the concept of a "digital continuum," bringing together HPC, AI, big data, IoT, cybersecurity, and mathematical modelling. TCI works to define specifications and recommendations for distributed systems that must integrate diverse computational and data-intensive technologies. Its outputs aim to inform EU research agendas and Horizon Europe missions, including climate adaptation, health, and digital twins. TCI is highly relevant for SPECTRUM, as it provides a cross-domain framework for integrating complex workflows that combine simulation, AI, and real-time data processing.

EUCloudEdgeIoT and the OpenContinuum project focus on building bridges across the cloud–edge–IoT ecosystem. OpenContinuum has produced a reference architecture structured around building blocks for security, trust, orchestration, data management, networking, monitoring, and AI/ML. A key contribution is its harmonised taxonomy, which aligns definitions across the cloud, edge, and IoT domains, enabling interoperability and reducing fragmentation. Its emphasis on embedding AI into all layers of the continuum, both for operational decision-making and for scientific analysis, offers a forward-looking model for how European research infrastructures could evolve.

IPCEI-CIS **(Important Projects of Common European Interest – Cloud Infrastructure and Services)** represents a large-scale industrial and governmental investment in Europe's cloud and digital services sector. The initiative aims to ensure European technological sovereignty by developing interoperable, federated cloud and edge infrastructures. Its reference architecture includes federated resource management, automated orchestration, and comprehensive security frameworks, designed for multi-provider and multi-tenant environments.

The EuroHPC Federation Platform **(EFP)**, launched in 2025, has the goal of making access to EuroHPC resources more seamless. At present, users must manage separate accounts, allocation processes, and software environments for each system. EFP seeks to unify this through federated identity and single

sign-on, consolidated allocation monitoring, and integrated workflow management across sites. Planned features include interactive SSH and web-based access (e.g. Jupyter, remote desktop), a federated software catalogue, and cross-system data transfer services. By Q1 2026, the platform aims to be in production, providing a more user-friendly interface to EuroHPC supercomputers, AI factories, and quantum systems. For researchers in HEP and RA, this is expected to simplify access to leadership-class resources while supporting heterogeneous and distributed workflows. The EFP is directly relevant to SPECTRUM's vision of a compute–data continuum, as it offers a practical foundation for federated access and resource interoperability, complementing the blueprint's emphasis on cross-facility workflows and harmonised user experience.