Federated HPC, Cloud and Data infrastructures

White Paper

Utz-Uwe Haus (HPE),
Sai Narasimhamurthy (Seagate),
Maria S. Perez (UPM),
Dirk Pleiter (KTH),
Andreas Wierse (SICOS)

11/04/2022

etp4hpc.eu
@etp4hpc

EUROPEAN TECHNOLOGY PLATFORM FOR HIGH PERFORMANCE COMPUTING

An increasing interest is observed in making a diversity of compute and storage resources, which are geographic spread, available in a federated manner. A common services layer can facilitate easier access, more elasticity as well as lower response times, and improved utilisation of the underlying resources. In this white paper, current trends are analysed both from an infrastructure provider as well as an end-user perspective. Here the focus is on federated e-infrastructures that among others include high-performance computing (HPC) systems as compute resources. Two initiatives, namely Fenix and GAIA-X, are presented as illustrative examples. Based on a more detailed exploration of selected topical areas related to federated e-infrastructures, various R&I challenges are identified and recommendations for further efforts formulated.

1. Introduction

Federating geographically spread HPC resources as well as storage resources has since long been an attractive vision. It matches, for instance, the needs of various distributed research communities, allows for more flexible access to more resources and access to a broader diversity of computing resources.[1] In this white paper, we discuss and describe trends that drive the realisation of federated e-infrastructures that include HPC resources. The documentation of selected initiatives show-case these trends. We furthermore explore different topical areas related to federated e-infrastructures in more detail to identify different R&I challenges, which are the basis for our recommendations. Throughout this white paper, we will focus on federated e-infrastructures that include HPC resources and that require tight integration and close proximity of computing and storage resources.

Many efforts have been invested in the past on federating HPC resources as a means of realising e-infrastructures that allow addressing large-scale scientific challenges. Early examples are the following projects that all started in the early 2000s: TeraGrid in the US [Catlett2005], DEISA in Europe [Gentzsch2011] and NAREGI in Japan [Matsuoka2005]. These projects had in common that they were based on Grid concepts and technologies, which had been developed mainly by academic organisations with rather limited commercial uptake. With the advent of commercial cloud providers that deployed federated compute and storage resources this situation fundamentally changed. Technologies required for federating compute and storage resources as well as services are now used at a much wider scale. Furthermore, they are driven by competing large-scale commercial entities, which increases the probability of different solutions being available.

One of the key assumptions made in this white paper is that infrastructures consisting of stand-alone HPC systems with attached high-performance storage resources will increasingly be replaced by e-infrastructures based HPC and cloud compute as well as storage resources that are made available through a set of federated services. In this context, federation means that the services are integrated in (one or more) access, identity and resource management mechanisms such that users can flexibly leverage different services, even if they are offered by different organisations at geographically different locations. These infrastructures should be realised without creating a risk for vendor lock-ins.

In the context of this white paper, it is important to note that the term "cloud" is used in different contexts. In the following, we will use this term to refer to technologies that are primarily used for realising public cloud infrastructures. We will not consider cloud business models although they may start to play an increasing role in future e-infrastructures that include HPC-based services.

[1] Examples are the bioinformatics community [ELIXIR] or various astronomy and particle physics communities [Swinbank2021].

2. Trends

In the past, the primary role of HPC centres was to operate one or more supercomputers and to provide access to this resource. In the future, we expect these centres to transform to providers of an *e-infrastructure services layer*, where the latter is based on different types of underlying compute and storage resources. While this will continue to include supercomputers, an increasing number of centres have also deployed cloud-type resources, i.e. on-premise private cloud instances. Most of the new EuroHPC pre-exascale and petascale systems do include such instances. These services started to become harmonised and federated across multiple sites, e.g. in the context of the Fenix initiative[2]. The trends sketched here have meanwhile become part of the future EuroHPC strategy, as EuroHPC added "HPC Federation and Services" as one of the five pillars of future activities [EU2020].

With these architectural changes, HPC centres respond to changing user needs as well as those of emerging new science and engineering domains, which do need HPC resources for their research. A first trend to highlight are efforts towards establishing domain-specific *platform services layers* that facilitate collaborative research within geographically dispersed communities. There are various opportunities for creating benefits through such higher-level services (see also [Netto2019]):

- Web Portals hiding complexity: The existence of user-friendly web-based interfaces can improve efficiency and increase the use of the infrastructures. Science Gateways [Gesing2018] are also used in HPC to increase the usability of modelling of data and simulations on top of complex computing infrastructures.
- The realisation of complex workflows: Realising workflows that combine simulations, big data processing and interactive steering can quickly become challenging due to the complexity resulting from the use of different, sometimes incompatible, tools and frameworks.
- Interactivity for enhanced execution steering: The choice of parameters and the tuning of the frameworks constitute a difficult task, which also impacts the application performance.
- Legacy applications transformation: The use of clouds and virtualization approaches enables the transformation of legacy applications for traditional computing infrastructures to a SaaS model (Software-as-a-Service).
- FAIR data management[3]: Platform services allow for integration with global data infrastructures and connect to data lakes that will grow to exabyte capacities.

Domain-specific platform services are being implemented among others by projects like the Human Brain Project[4] or different ESFRIs like those organised in the European Science Cluster of Astronomy & Particle physics ESFRI research infrastructures (ESCAPE)[5] [Dona2021] or the ESFRI for life-science information Elixir[6].

There are several benefits that can be achieved by federating e-infrastructure services. Firstly, research communities can deploy domain-specific services without becoming dependent on a single resources provider. Secondly, enabling data sharing and exploiting data locality become easier in a federated infrastructure. Data does not have to be collected in a single location but can remain close to where it has been produced while compute resources can be allocated near the data. Furthermore, federated infrastructures can help to improve the availability of services by geo-replication of services and data. If one site becomes unavailable then both services and data remain accessible at other sites. Finally, federated infrastructures can provide more flexibility to serve a diversity of communities as more resources are available for allocation and a larger variety of services can be offered through different resource providers.

[2] https://www.fenix-ri.eu/
[3] https://www.go-fair.org/fair-principles
[4] See, e.g., EBRAINS Simulation services offering (https://ebrains.eu/services#category2)
[5] https://projectescape.eu/
[6] https://elixir-europe.org/

3. Selected initiatives

There are a large number of initiatives throughout Europe towards the federation of e-infrastructure services. We introduce in the following two initiatives as illustrative examples that have a connection to HPC and involve various ETP4PHC members.

3.1 Fenix

Fenix [Fenix2021, Alam2020] is a collaboration of HPC centres working on the harmonisation and federation of their offerings of e-infrastructure services with the goal of supporting a variety of science and engineering communities. This service portfolio's distinguishing characteristic is that different types of data repositories, scalable supercomputing systems, and private cloud instances are in close proximity and thus well connected and integrated. The different sites that act as e-infrastructure services providers are interconnected via a high-speed network. A key feature of the Fenix approach is the separation of concern with HPC centres being in the role of providing a set of federated e-infrastructure services that support user communities that deploy their domain-specific services. A schematic view is provided in *Figure 1*.

Fenix has the ambition of serving in a sustainable manner relevant science and engineering domains that strongly benefit from diverse e-infrastructure services for their collaborative research. For being able to scale to a larger number of such domains, Fenix focuses on a consolidated portfolio of services. To stay aligned with the needs of current and upcoming science and engineering domains, Fenix governance foresees a representation of these domains such that they can drive the evolution of the e-infrastructure services portfolio.

Fenix members leverage national, European and international funding programs to realise the compute, storage and network resources sustaining the e-infrastructure services. The coherence of the approach is ensured through a clear governance model and close technical collaboration at various levels. Furthermore, the Fenix partners share the responsibility for the operation of federation-level services and the integration at each of the sites.

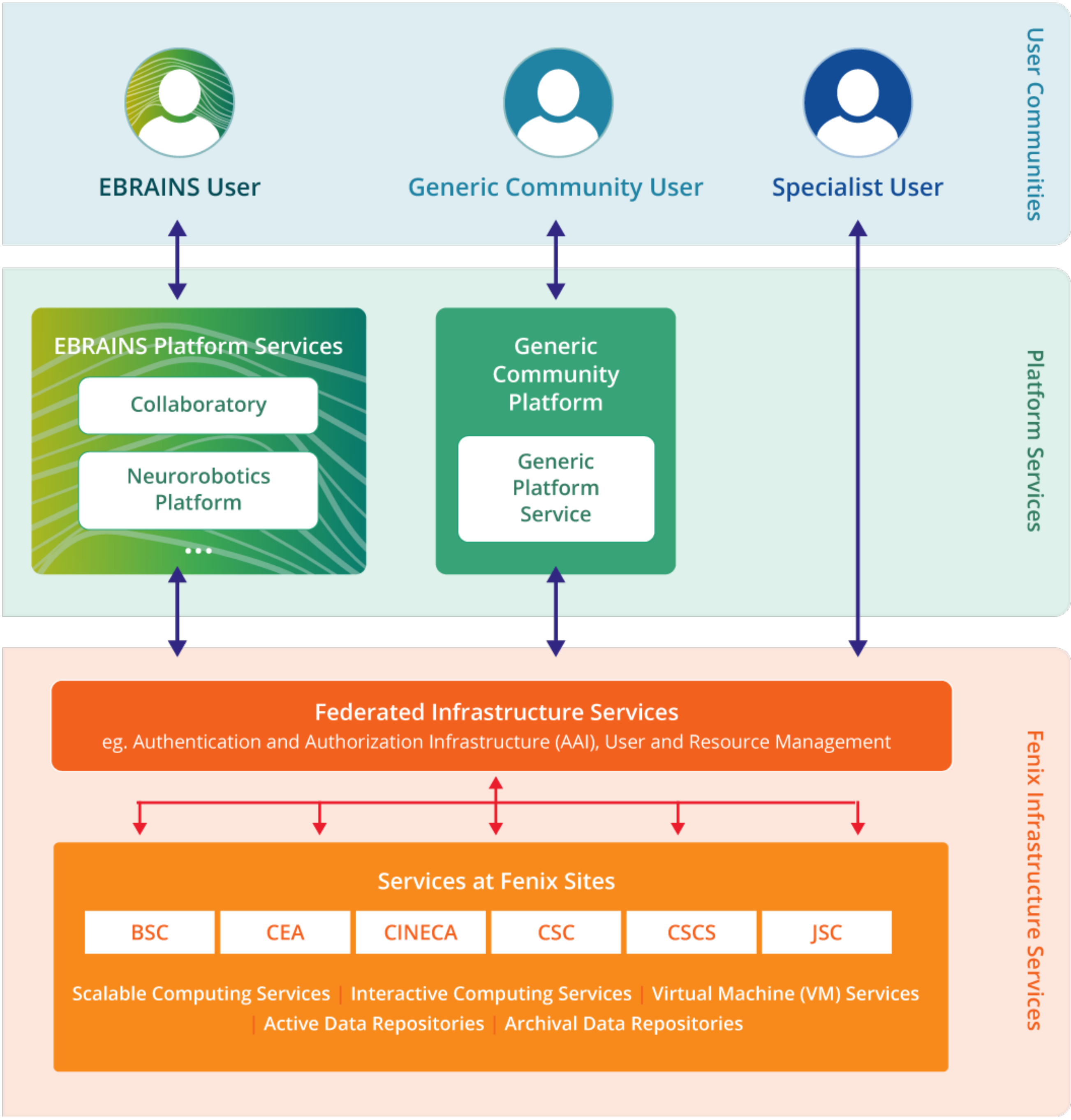

Figure 1 - Schematic overview of Fenix's layered approach for support users from different science and engineering domains (source: Fenix)

3.2 Gaia-X

The backdrop of Gaia-X can be traced back to late 2019. There had been a general concern in Europe for many years that big global hyperscalers (like Amazon or Alibaba) outside of Europe handle the data of European citizens and organisations. Hence, there existed a political will to create something in Europe that breaks European dependency on these big non-European hyperscalers. Thus arose Gaia-X, which was a Franco-German initiative to create a European data infrastructure and data economy.

In Europe, there are many challenges to creating an indigenous data economy. Firstly, there are decentralised data processing locations throughout Europe, which are not consolidated. These locations all use their own stacks. Interoperability and flow of data between these locations are not seamless. There is an absence of widely accepted APIs to share data and insufficient clarity about applicable jurisdiction. Also, there are sector-specific data spaces (e.g. automotive, financial, pharma) and sharing data across sectors is extremely hard.

The Gaia-X solution envisioned a sharable data ecosystem between various data spaces all relying on a federated infrastructure ecosystem to store data. The federated infrastructure ecosystem consists of existing network and interconnection providers (e.g., telcos), Cloud Service Providers (CSPs), sector-specific clouds, edge locations and HPC. The data spaces and the infrastructure ecosystem are all orchestrated by the Gaia-X Federation Services, which deals with issues of identity and trust management, provision and maintenance of a federated catalogue of infrastructure providers, and compliance management. It is envisioned that smart third party services (AI, IoT, etc) could be built on top of these data spaces. The key European data spaces identified are agriculture, energy, health, industry 4.0, mobility, public sector, smart living and finance.

Figure 2 shows the Gaia-X architecture and the role of HPC as envisioned there. More details can be found in the Gaia-X architecture document [Gaia-X2021].

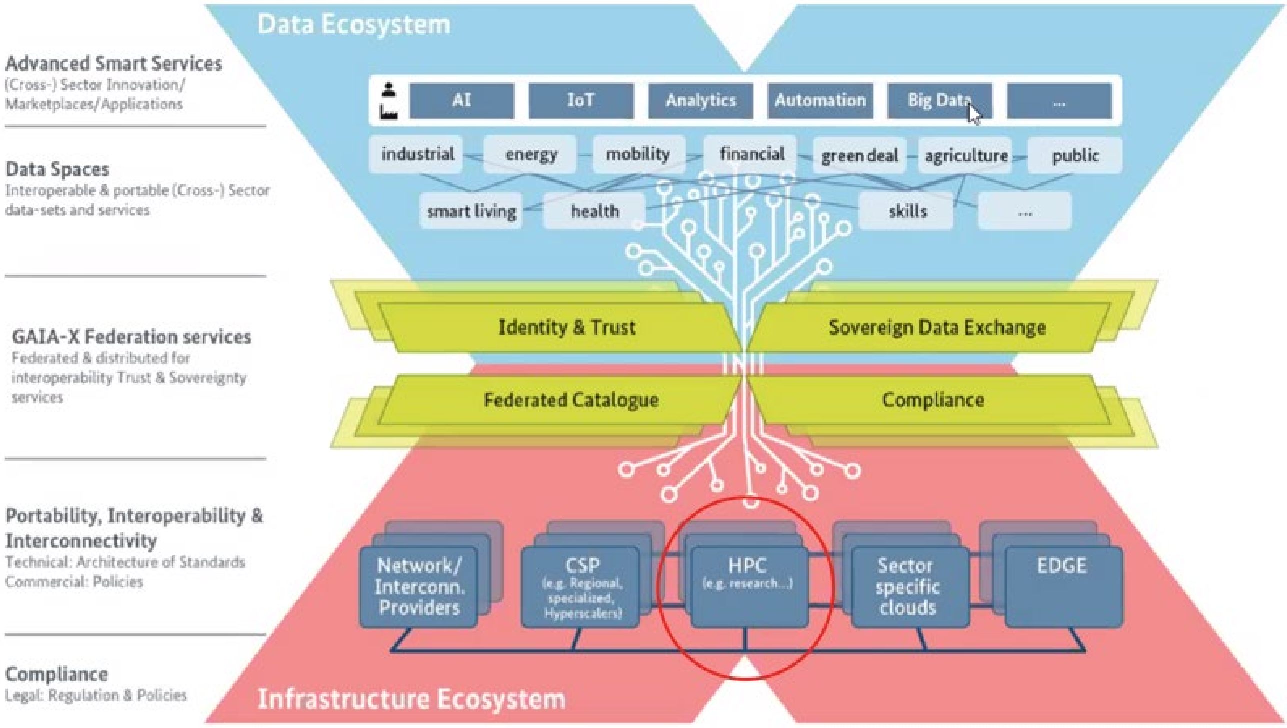

Figure 2 - Overview of Gaia-X (gaia-x.eu)

Some of the key considerations of the Gaia-X Infrastructure ecosystem are:

1. Establishing portability and interoperability amongst the different infrastructure providers (HPC, edge, CSPs, etc);
2. Providing the ability for finding, combining and connecting services from participating providers.

At the time of writing this white paper, Gaia-X has been fully set up as an AISBL organisation with a fully functioning organisational structure - Policy & Rules Committee, Data Spaces Business Committee and a Technical Committee consisting of various working groups looking into different aspects from the perspective of Users, Providers, Architecture, etc. Various Gaia-X hackathons are being organised to develop proofs-of-concepts supporting Gaia-X and the first version of the fully-functioning Federated Services is expected to become available in 2022.

HPC as a part of a Gaia-X infrastructure ecosystem is a subject of current discussions. HPC sites in Europe may become resource and service providers within Gaia-X. Some of the issues that need to be addressed are how the HPC services describe themselves within Gaia-X, what are the possible economic models for their engagement, how they work vis-a-vis other initiatives such as the European Open Science Cloud, etc. One possibility is the focus on industrial usage of HPC, since Gaia-X seems to be heading in the direction of being Europe’s “Industrial Federated Cloud”, considering very strong participation from the Industrial actors. The focus on HPC within Gaia-X creates the opportunity for making HPC services available to a wider user community as well as to drive the development of solutions for federated services based on HPC resources.

4. Topical areas

In this section a few topics are considered that are relevant for realising future federated e-infrastructures including HPC resources.

4.1 Authentication and authorisation infrastructure

Federation of e-infrastructure services requires integration in a common Authentication and Authorisation Infrastructures (AAI) such that a user with a single identity as well as one (or few) credentials can authenticate with different services.

A typical AAI architecture for federated e-infrastructures foresees, on the one hand, one or more identity providers and on the other hand a set of service providers, which are geographically distributed in case of federated e-infrastructures considered here. Different initiatives, including Fenix or EOSC [EOSC2021], foresee such an architecture as they follow the architecture proposed by the AARC project [AARC2019] with important elements being realised by eduGAIN[7]. One key challenge for realising such an architecture is the establishment of suitable trust relations between identity providers and service providers. This is particularly relevant for sensitive services, which includes any service that provides direct access to HPC resources. For such services, a high level of assurance is required, while other services (e.g. web portals) may require only a low level.

For a federated e-infrastructure, authentication mechanisms need to be harmonised and support multiple authentication mechanisms to facilitate user-friendly access to different services. While authentication to HPC resources today typically relies on authentication using SSH, different technologies like OpenID Connect (OIDC)[8] are used for cloud-type services. At the same time, security will have to be pushed to a higher level, e.g. by replacing today's dominant single-factor by multi-factor authentication mechanisms. In the context of a federated e-infrastructure, there are additional security concerns as a single compromised identity would put more service providers at risk. Today many services are accessible using a single-factor for authentication. There is an emerging consensus that soon multi-factor authentication will be required, which is likely to become mandatory due to European or local regulations.

Also, authorisation mechanisms need to be established at the level of a set of federated services. While this may generically be solved through attribute services as foreseen in the AARC architecture, in practice interoperability of different solutions is typically at best at an early stage.

4.2 Integration of HPC- and Cloud-based compute and data services

For multiple reasons, services based on HPC and cloud instances are still difficult to integrate. While direct access to HPC resources requires strict control, cloud instances are typically more openly accessible. Another reason is the diversity of how HPC resources are operated at different sites. This makes, for instance, the deployment of web-based portal services that facilitate the spawning of HPC jobs difficult and tedious due to the required customisation. Such kind of functionality is expected to become increasingly important in the context of community-specific platform services layers (see Section 2), the use of HPC for urgent decision making (see [Marazakis2022]) or the realisation of use cases identified by the Transcontinuum Initiative.[9]

Similarly, federating HPC-based services and data services remains challenging as integration is still often proprietary. An increasing number of data platforms are created and realised through lake concepts, but implementing workflows that consume data available on such platforms or publish data produced on the HPC system often requires manual steps. They can often not be automated as handling of the necessary credentials is not supported.

---

[7] https://edugain.org/
[8] https://openid.net/connect/
[9] https://www.etp4hpc.eu/tci-use-cases.html

4.3 Resource management, allocation and accounting

HPC and cloud infrastructures tended to adopt different approaches to resource management. In the area of HPC, resource managers compromise on response times to maximise resource utilisation without oversubscribing compute resources; they are allowed to significantly delay start-up of jobs. In the area of cloud, the focus is rather on short response times and elastic use of the available hardware resources. Another difference is that HPC resources are typically assumed to be concurrently used by a smaller number of long-running large-scale jobs (in the extreme case a single job), while cloud computing resources may be used by a very large number of small microservices, which may run for only short periods of time, so that set-up and tear-down latencies become critical. Both HPC and cloud computing resource management is becoming more complex when underlying hardware resources become more heterogeneous (for instance, when part of the servers host compute accelerators such as GPUs).

For an e-infrastructure comprising HPC and cloud resources this means that different types of resource managers will co-exist. Additionally, in a federated e-infrastructure coordination between resource managers at different sites may be required to ensure timely availability of resources. This is an area ripe for further collaboration between the developers and providers of services based on HPC and Cloud resources and technologies. Further research and innovation is needed here, for example on how to orchestrate different types of resource managers to optimally use the available HPC or cloud infrastructure, while supporting co-allocation of different types of resources. Innovations in this area will also help in the seamless transition to HPC cloud bursting - where HPC centres can leverage cloud services on demand when existing HPC infrastructure has to deal with demand increases that cannot be met locally.

Another challenge in the context of federated e-infrastructures is centralised allocation of resources and accounting of used resources. Such a capability would on the one hand allow to allocate resources that are made available by different organisations to single projects such that, for instance, a project could have different types of computing resources at different locations. On the other hand, members of such projects would be able to monitor consumption of these resources at a central location without going through a process of collecting this information from different resource providers using different proprietary interfaces. Realising such capabilities would, in particular, require standardised and protected interfaces and mechanisms for distributing information on resource allocation and consumption. One example for such a centralised resource allocation and accounting service is the Fenix FURMS service.[10]

4.4 Trust, security and data compliance

Federated e-infrastructures integrate geographically distributed resources that are operated in different security domains such that data will be transferred over organisational boundaries. To use such e-infrastructures for workflows, for which confidentiality and security are particularly critical (e.g. to protect personal data or trade secrets), requires reconsideration of existing security measures, a more active security management targeting harmonised security levels at different sites as well as suitable mechanisms for establishing trust. While some aspects need to be addressed at a policy level others require suitable technical solutions.

As an example, we consider the need for an end-to-end secure setup for workflow execution. Secure and trustworthy execution of HPC workloads is in its infancy while cloud system operations often rely on virtualisation to provide this, from node attestation to verifying workload attestation. The HPC community efficiency concerns prohibit such costly isolation procedures or the use of trusted execution environments (TEE).[11] This will become more challenging within a federated infrastructure with multiple and heterogeneous sets of computing resources with different hardware support for TEEs. Future middleware solutions, in concert with the scheduling system, will need to address this challenge. Furthermore, trustable deployment of workload and other software components must be possible. These may be deployed using containers within an e-infrastructure supporting available security guidelines [Souppaya2017] and container encryption. Finally, the trust may be enhanced by putting mechanisms in place for auditing distributed workflow execution. These can be based on distributed solutions for generating audit trails using blockchain technologies (see, e.g., [Al-Mamun2019]).

---

[10] https://github.com/unity-idm/furms

[11] See [Akram2021,Peisert2021] for recent studies on trusted environments in the context of scientific computing workloads.

Another example concerns data compliance. While data is transferred over organisational and possibly state boundaries, the complexity of maintaining compliance with regulatory, organisational, or contractual data handling requirements becomes more complex. Data management solutions may support enforcing compliance, for instance by tagging data repositories and services depending on the met requirements (see, e.g., [Henze2017]).

4.5 Distributed infrastructure monitoring

One key challenge for federated infrastructures is to monitor (and operate) them in such a way that in particular end-user services function seamlessly and with very high availability. Any malfunction or performance degradation needs to be identified in a central manner within a short period of time. Continuous testing and monitoring mechanisms must be organised in such a manner that difficult to understand error patterns, which are the result of a complex interplay between different services, can be identified.[12] Today, realising monitoring of distributed infrastructure components still requires a significant amount of customisation due to a lack of interfaces for collecting the necessary information with suitable access control.

Monitoring is expected to become more important as distributed infrastructures need to be able to support workflows with particular service quality demands, for instance in terms of service availability. A prominent example is numerical weather prediction, where particular calculations need to have been completed within pre-defined (short) periods of time as they otherwise would become obsolete. Today typically dedicated resources are provided for weather services, but in future also general-purpose resources provided through, e.g., the EuroHPC Joint Undertaking are expected to be used [EU2020]. Monitoring mechanisms therefore need to support compliance with service-level agreements.

4.6 Harmonisation of service provisioning

To ensure coherence within a federate e-infrastructure, harmonisation of the service provisioning is an important aspect. This can be supported by promoting the use of automatized service deployment using tools like Ansible[13], Puppet[14], or Terraform[15]. Sharing the relevant configuration files will not only improve harmonisation but also improve commoditization of the services.

Another area concerns the harmonisation of software environments within a distributed e-infrastructure. Different strategies started to be explored and implemented. The most popular approach is the use of containers, an OS-level virtualisation technique that facilitates deployment of customised environments within different host environments. It has become a standard technology in the cloud environment and is meanwhile commonly supported on HPC systems. To avoid customisation for individual users and user groups, tools like EasyBuild[16] and Spack[17] can be used as they allow sharing recipes for software installation that can be used on rather different systems. Finally, there are efforts for creating unified software environments that can be distributed to different systems of different types (see, for instance, [Boissonneault2016].

4.7 Usability by end-users

Federation of different types of e-infrastructure services open many new opportunities for applications that require using a combination of services based on HPC, cloud computing and various storage resources. The lack of usability is one of the main drawbacks that both applications and tool developers are currently facing. Improving usability deserves significant efforts in order to leverage the opportunities offered by these e-infrastructures and to increase their effective use.

---

[12] As an example we refer here to the ESCAPE data lake testing infrastructures [Dona2021].
[13] https://www.ansible.com
[14] https://www.puppet.com
[15] https://www.terraform.io
[16] https://docs.easybuild.io/en/latest/
[17] https://spack.readthedocs.io/en/latest/

With regards to new research and innovation efforts in terms of usability, the following are some of the most promising approaches:

1. Establish DevOps for federated e-infrastructures: DevOps has been introduced as a term to refer to sets of common practices that bring together software development and infrastructure operations [Desbiens2021]. Such concepts have been applied in other areas to facilitate the process of development while ensuring the deployed code fits with the characteristics of infrastructure. In the area of AI-based services, the term AIOps was introduced to refer to engineers capable of efficiently and effectively building and operating online services that use AI and techniques [Dang2019]. Translating this philosophy to federated e-infrastructures including HPC will contribute to the usability and improvement of solutions tailored to these e-infrastructures, and therefore help to increase their popularity.
2. Services and resource selection support: Within distributed e-infrastructures the process of deciding on the choice of services and resources can become very complex when taking into account, for instance, cloud costs and benefits trade-offs as well as data movement policies to increase data locality.
3. Intelligent Automation for knowledge and service work: The term Intelligent Automation has been introduced to describe application of AI in ways that can learn, adapt and improve over time to automate tasks that as of today are performed by humans [Coombs2020]. Examples are decision support and expert systems or recommendation agents. This approach could benefit from the aforementioned AIOps.
4. Creation of value-added cloud services integrated with HPC: A usable approach is characterised by providing value to the consumers. The virtualization of HPC resources and its deployment in a cloud-based approach, on the one hand, and the extension of data centre capabilities with value-added services, on the other hand, will provide usability, scalability, elasticity and will decrease the total cost of ownership (TCO).

5. Recommendations for R&I efforts

On the basis of the analysis of the different topical areas the following recommendations for further research and innovation efforts by ETP4HPC members, in particular in the context of R&I actions supported by the EuroHPC Joint Undertaking:

1. The integration of HPC systems and HPC-based services within the AAI of a federated e-infrastructure should become a commodity while at the same time different and sufficiently strong authentication mechanisms should be supported, e.g. through multi-factor authentication.
2. The integration of different types of services based on secured HPC and data resources and other services that can be made more openly available, e.g. web-based portal services, needs to be improved. Specific examples are web-based APIs to resource managers, support of data management involving HPC systems and data platforms.
3. Improve support of centralised resource allocation and resource consumption monitoring within a federated e-infrastructure with different types of resources provided by different organisations.
4. Seek collaboration between HPC and cloud communities for research and innovation on orchestration of different types of resource management systems within a federated e-infrastructure.
5. Improve support for end-to-end secure setup for workflow execution, mechanisms for trustable deployment of workload and other software components as well as solutions that help to enforce data compliance.
6. Establish standardised interfaces for collecting monitoring information with suitable access control within a distributed environment.
7. Promote usability by end-users by establishing DevOps for federated e-infrastructures, developing services and resource selection support tools and exploring Intelligent Automation for knowledge and service work.

6. Summary and conclusions

In this white paper, we discussed general trends towards integration of HPC systems within federated e-infrastructures and presented two initiatives that are working on establishing such e-infrastructures. While such e-infrastructures have been established and are in use, there are still many areas where research and innovation efforts are needed for commoditization, improved capabilities and usability. We explored some of these areas in more detail and made a number of specific recommendations for research and innovation efforts in the context of ETP4HPC.

References

[AARC2019] Nicolas Liampotis et al., “AARC Blueprint Architecture 2019”, 2019 (DOI: 10.5281/zenodo.3672785)

[Akram2021] Ayaz Akram et al., “Performance Analysis of Scientific Computing Workloads on General Purpose TEEs”, IPDPS, 2021 (DOI: 10.1109/IPDPS49936.2021.00115)

[Al-Mamun2019] Abdullah Al-Mamun et al., “HPChain: An MPI-Based Blockchain Framework for Data Fidelity in High-Performance Computing Systems”, Supercomputing, 2019 (https://sc19.supercomputing.org/proceedings/tech_poster/poster_files/rpost106s2-file3.pdf)

[Alam2020] Sadaf Alam et al., “Archival Data Repository Services to Enable HPC and Cloud Workflows in a Federated Research e-Infrastructure”, IEEE/ACM International Workshop on Interoperability of Supercomputing and Cloud Technologies, 2020 (DOI: 10.1109/SuperCompCloud51944.2020.00012)

[Boissonneault2016] Maxime Boissonneault et al., “Providing a Unified Software Environment for Canada’s National Advanced Computing Centers”, PEARC’19, 2019 (DOI 10.1145/3332186.3332210)

[Catlett2005] Charles Catlett et al., “TeraGrid: Analysis of Organization, System Architecture, and Middleware Enabling New Types of Applications”, in: L. Grandinetti (Ed.), “High-Performance Computing and Grids in Action”, IOS Press, 2008.

[Coombs2020] C. Coombs, D. Hislop, S. K. Taneva and S. Barnard, “The strategic impacts of Intelligent Automation for knowledge and service work: An interdisciplinary review”, The Journal of Strategic Information Systems, 2020 (DOI: 10.1016/j.jsis.2020.101600)

[Dang2019] Y. Dang, Q. Lin and P. Huang, "AIOps: Real-World Challenges and Research Innovations”, IEEE/ACM 41st International Conference on Software Engineering, 2019 (DOI: 10.1109/ICSE-Companion.2019.00023)

[Desbiens2021] Frédéric Desbiens, “EdgeOps for IoT: A Vision For Edge”, Connected World, 2021 (https://connectedworld.com/edgeops-for-iot-a-vision-for-edge/)

[Dona2021] Rizart Dona, Riccardo Di Maria (ESCAPE project), “The ESCAPE Data Lake: The machinery behind testing, monitoring and supporting a unified federated storage infrastructure of the exabyte-scale”, CHEP 2021, EPJ Web of Conferences, 2021 (DOI: 10.1051/epjconf/202125102060)

[ELIXIR] ELIXIR Compute Platform, https://elixir-europe.org/platforms/compute (access 27.03.2022).

[EOSC2021] European Commission, Directorate-General for Research and Innovation, Vaghetti, D., Kanellopoulos, C., Johansson, L., et al., “EOSC Authentication and Authorization Infrastructure (AAI): report from the EOSC Executive Board Working Group (WG) Architecture AAI Task Force (TF)”, Publications Office, 2021 (https://data.europa.eu/doi/10.2777/8702)

[EU2020] European Commission, “Equipping Europe for world-class High-Performance Computing in the next decade”, SWD(2020) 179 final, https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52020SC0179&rid=9.

[Fenix2021] Fenix, “Fenix Strategy Document”, 2021 (https://fenix-ri.eu/sites/default/files/public/file-uploads/Fenix%20Strategy%20Document%20202111111-public.pdf)

[Gaia-X2021] Gaia-X, “Gaia-X Architecture Document”, 21.09 release, 2021 (https://www.gaia-x.eu/sites/default/files/2021-10/Gaia-X_Architecture_Document_2109.pdf)

[Gentzsch2011] Wolfgang Gentzsch et al., “DEISA - Distributed European Infrastructure for Supercomputing Applications”, J Grid Computing, 2011 (DOI: 10.1007/s10723-011-9183-2)

[Gesing2018] S. Gesing, “Science Gateways in HPC: Usability Meets Efficiency and Effectiveness” In: Modelling and Simulation in HPC and Cloud Systems, Book Series "Studies in Big Data" Print ISBN: 978-3-319-73766-9, Electronic ISBN: 978-3-319-73767-6, Springer International Publishing, 2018

[Henze2017] Martin Henze et al., “Practical Data Compliance for Cloud Storage”, IEEE International Conference on Cloud Engineering (IC2E), 2017 (DOI: 10.1109/IC2E.2017.32)

[Matsuoka2005] Satoshi Matsuoka et al., “Japanese Computational Grid Research Project: NAREGI”, Proceedings of the IEEE, Vol. 93, Issue 3, 2005 (DOI: 10.1109/JPROC.2004.842748)

[Marazakis2022] Manolis Marazakis et al., “HPC for urgent decision making”, ETP4HPC White Paper, 2022 (DOI: 10.5281/zenodo.6107362)

[Netto2019] Marco A. S. Netto et al., “HPC cloud for scientific and business applications: taxonomy, vision, and research challenges”, ACM Computing Surveys (CSUR) 51 (1), 1-29, 2019 (DOI: 10.1145/3150224)

[Peisert2021] Sean Peisert, “Security Trustworthy Scientific Computing”, Communications of the ACM, 2021 (DOI: 10.1145/3457191)

[Souppaya2017] Murugiah Souppaya, John Morello, Karen Scarfone, “Application Container Security Guide”, NIST Special Publication 800-190, 2017 (DOI: 10.6028/NIST.SP.800-190)

[Swinbank2021] John D. Swinbank et al. (ESCAPE), “D5.3 - Performance Assessment of Initial Science Platform Prototype”, 2021 (https://projectescape.eu/sites/default/files/ESCAPE-D5_3.pdf)

Glossary and list of acronyms

AAI	Authentication and authorization infrastructure
AI	Artificial intelligence
API	Application programming interface
CSP	Cloud service providers
DevOps	A practice that combines software development (Dev) and IT operations (Ops)
EOSC	European Open Science Cloud (https://eosc-portal.eu/)
ESFRI	European Strategy Forum on Research Infrastructures
HPC	High-performance computing
IoT	Internet-of-things
Level of assurance	Degree of confidence in the claimed identity of a person
ML	Machine learning
OIDC	OpenID Connect (https://openid.net/connect/)
OS	Operating System
TEE	Trusted execution environment

Authors:

Utz-Uwe Haus is Head of HPE HPC/AI EMEA Research Lab, and was technical coordinator of the MAESTRO project.

Sai Narasimhamurthy leads European R&D for Seagate Systems and led the SAGE & Sage2 storage projects.

Maria S. Perez is a Full Professor at Universidad Politécnica de Madrid.

Dirk Pleiter is a Professor of High Performance Computing at the Kungliga Tekniska högskolan Royal Institute of Technology and Director of the PDC Center for High Performance Computing (KTH - Sweden).

Andreas Wierse is the managing director of SICOS BW that supports SMEs in the use of high performance computing.

Cite as: Utz-Uwe Haus et al., « Federated HPC, Cloud and Data infrastructures», ETP4HPC White Paper, 2022, doi 10.5281/zenodo.6451288.

DOI: 10.5281/zenodo.6451288