A Vision for a European e-Infrastructure for the 21st Century

Executive Summary

Over the past decade Europe has developed world-leading expertise in building and operating very large scale federated and distributed e-Infrastructures, supporting unprecedented scales of international collaboration in science, both within and across disciplines. We have the opportunity now to capitalize on that investment and experience, to build the next generation infrastructure to enable innovation and opportunities for European science and education, industry and entrepreneurs.

We are now in a period of explosive data growth. The foundations for handling the "Data Tsunami" or "Big Data" have been laid in the last 20 years as we have moved from simple commodity computing ("Farms"), to commodity distributed computing ("Grid") and then commodity computing services ("Cloud"). These have prepared the ground for handling the large amounts of data being produced today. The era of "Data Intensive Science" has begun.

To address these challenges for the diverse, emerging "*long tail of science*" conducted by researchers that do not have access to significant in-house computing resources and skills, we propose creating a common platform for the future that builds on the experience of the last decade and is flexible enough to adapt to technological and service innovations. Such a platform must provide the underlying layers of common services, but must be adaptable to the very different and evolving needs of the research communities. A key feature should be that established services be operated by European industry, while development of new services may be publicly funded. The proposal has 3 distinct layers of services:

1. European and international networks; services for identity management and federation across all European research and education institutions and integrated with other regions of the world;
2. A small number of facilities to provide cloud and data services of general and widespread usage.
3. Software services and tools to provide value-added abilities to the research communities, in a managed repository:
    a. The tools to provide those research communities that have access to large sets of resources the ability to federate and integrate those resources and to operate them for their community, potentially sharing with other communities;
    b. Tools to help build applications: e.g. tools to manage data, storage, workflows, visualisation and analysis libraries, etc.
    c. Tools and services to allow researchers to integrate everyday activities with the e-Infrastructure: collaborative tools and services; office

automation, negotiated licensing agreements etc. Services would be operated by industry or on the facilities in layer 2 above;

d. Tools to help research communities engage the general public as citizen scientists.

These layers would be supplemented by investment in application software in order to build and share expertise in ensuring that applications are capable of exploiting evolving computing architectures.

The expectation is that a continuum of financial models is appropriate ranging from sponsored resources for peer-reviewed scientific cases to communities who would pay for the services they receive, thus the services they receive must be appropriate and provide a clear value. The governance of the platform would be created by representation from the user communities.

Contents

This document has been prepared by the IT department of CERN on behalf of the EIROforum IT working group.

Introduction

Looking forward over the coming 10-15 years, there are exciting challenges ahead to capture, manage, and process the vast amounts of data likely to be generated, not only by the established fundamental research domains but in a growing range of scientific disciplines, large and small. This new frontier in computation will be driven by the needs of data-driven science, simulation, modelling and statistical analysis in areas from climate change to life sciences, art and linguistics. All will see incredible growth and accelerated breakthroughs due to unprecedented access to data and the computational ability to process it.

Europe must preserve its intellectual capital and provide the opportunity for it to be nurtured, developed and to grow. Major advances in technology that have taken the world by storm, from Linux to the World Wide Web, have often been conceived in Europe but ultimately exploited elsewhere. This is a loss to Europe, in terms of skills, employment and business.

We must take the step to make unprecedented scales of IT resources available to the next generation of emerging scientists, researchers and entrepreneurs, nurturing them from education to start-up activities and then sustainable businesses or research communities. As well as serving the direct needs of computing, the opportunity to innovate and explore new technologies is essential. The correct environment for innovation will allow many ideas to be tested, and explored, which will lead to the truly unexpected and ground-breaking discoveries and inventions that will shape this century.

Over the last decade, driven with sustained funding from the EC, the e-Infrastructure landscape across Europe has grown from regional prototypes to a set of pan-European production resources. But to go forward much more coordinated effort is needed. Today's efforts leave gaps in the overall strategy, and suffer in part from inadequate funding by the stakeholders.

CERN, in collaboration with the EC, national funding agencies and the High Energy Physics community has successfully built, and today operates, the world's largest scientific e-infrastructure. This worldwide infrastructure is in daily operation and has been used to produce results from the huge volumes of data delivered by the LHC and its detectors. The development of this distributed grid federating resources around the world took close to 10 years from conception through to production use at the necessary scale, and required novel developments in terms of physical infrastructure, middleware, application software, and policy development. This experience also highlights a key aspect of the future e-infrastructure model if it is to become *the* infrastructure of choice for the European Research Area: there must a long-term commitment by all the stakeholders to make the e-Infrastructure the means by which they will provide/use production IT services. The future research infrastructures currently in construction, such as FAIR, XFEL, ELIXIR, SKA, ITER and upgrades to ILL and ESRF, need to be convinced that the e-Infrastructure will exist and continue to

evolve throughout their construction and operation phases if they are to take the risk and invest in its creation and exploitation.

In considering the way forward, it is important that we foresee an infrastructure that supports all of the scientific and academic needs of the European community, including the *"long tail of science"* conducted by researchers that do not have access to significant in-house computing resources and skills. Consequently this should not be thought of as a one-size-fits-all solution. Rather a broad but coherent set of services and tools which must be available to allow the specific needs of each community to be met. This common platform should also be able to act as the incubator for new businesses and scientific activities. It is essential that European industry engage with the scientific community in building and providing such services, but it is also important that the user community have a strong voice in the governance. This view has been documented in a recent Response to EC (DG CNECT) Paper "Research Data e-infrastructures: Framework for Action in H2020" produced by the EIROforum IT Working Group.

Proposal for an e-Infrastructure

While the grid model has been extremely successful for High Energy Physics and similar high-throughput computing applications (such as astrophysics), it is not suited for many other sciences, which have very different requirements. Technological advances are continuous, and so during the time of the development of the grid, distributed computing and the underlying technologies have advanced significantly, in academia and have been adopted by many business sectors. Cloud computing technologies, and the huge increase in available networking capabilities are leading examples. The growing computing needs of sciences, in particular those that have never before needed large scale computing, will benefit from many of those advances. Thus it is vital for the future needs of scientific e-infrastructures, that a model be adopted encompassing a wide spectrum of facilities and tools that can be of direct benefit to a range of science and research use cases. Rather than build such a structure as a single integrated e-infrastructure (like the grid) it will be far more advantageous to provide a set of collaborating core infrastructure services, and a variety of facilities, together with a broad set of easily adaptable tools. This will allow the research communities to select the services or tools that they require, and only those, without additional complexity. One of the lessons from the grid experience is that unnecessary complexity should be avoided in the infrastructure layers.

There are many existing efforts within Europe that can be drawn on to fulfil a vision for the future. These "pathfinder" initiatives have prototyped many aspects of what will be needed in the future. This includes much of the work in the grid projects, but also projects such as EUDAT, CRISP, Helix Nebula, OpenAIRE, thematic data projects, such as Transplant and many others.

In order that such an effort be sustainable and permit maximum flexibility across domains, as well as being able to fulfil the goals of working together with industry and key global players, it is essential that the future service infrastructure and tools be fully based on open standards, open software, and provide open access to the data.

This integrated model representing a common platform for the future must have the following key components:

A set of basic infrastructure services:

- The core network; Building on today's GEANT/NREN, and commercially operated networks to provide excellent connectivity and operational services to all scientific institutions, and fully extending to countries on the edges of Europe, as well as ensuring the international and global connectivity required by today's sciences;
- Federated identity management services, allowing existing identities of researchers to be used across the full set of e-Infrastructure services. Bearing in mind the substantial achievements of the ORCID project, it must support persistent digital identifiers that uniquely distinguish every researcher throughout their entire career, providing integration in key research workflows such as manuscript and grant submission, supports automated linkages between professional activities and ensuring the researcher's work is recognised. The FIM4R document[1] produced by representatives from a range of research disciplines provides detailed requirements for such services.

Cloud services

A (small number of) publicly provided research community cloud facilities that can be used for applications that require the instantiation of a few long-lived services, or access to compute or storage resources for a relatively short time. Such a cloud resource could be outsourced to commercial providers, or be the product of a public-private partnership. The Helix Nebula project provides an example for what may be involved in putting this in place.

Data Facilities

A general data storage facility (for example public science archives). Not only would such a resource be of immense value to data producers, providing a sustainable, dependable and accessible archive; but also would provide an unrivalled opportunity for data sharing between sciences with the integration of different types and sources of data. The EUDAT project is demonstrating some of the candidate technologies in this area. These facilities would provide open access to the data, and would be a focus for data preservation activities to ensure the long-term guardianship of the data. The facilities would also provide persistent identifiers for data objects at an appropriate granularity, as well as metadata services. They would allow secure data sharing between sciences capable of supporting certification requirements of Data Access Committees. As such the e-infrastructure will offer a knowledgebase consisting of a collection of data, organizational methods, standards, analysis tools, and interfaces representing a body of knowledge.

[1] Federated Identity Management for Research Collaborations, http://cds.cern.ch/record/1442597

Distributed infrastructure

A set of high-level software services that allow research communities to implement a federated and distributed computing infrastructure in order to integrate resources often explicitly provided for those applications. These are typically useful where the computing and storage requirements are large, where there is a need to collaborate, and where the occupation of the resources is very high. These services would be a generalisation of today's grid services, but should focus on the move to more open and standard implementations, and may benefit directly from cloud implementations. By democratizing access to data and computational resources, the services will enable any laboratory or project, regardless of size, to participate in a transformative community-wide effort for advancing science and accelerating the pace toward its exploitation. Thus, the e-infrastructures services will facilitate building a broader scientific community that will contribute to fundamental science within the European Research Area.

Software services and tools

1. A set of software services that allow researchers to integrate e-infrastructure with their everyday activities and personal devices, for example a "dropbox" functionality, office automation and collaborative tools and services. Many of these would be hosted on the infrastructures described above. Today there are many tools available but most are not widely known or used. This action would also provide a means by which software licenses (which today represent a significant and rapidly growing cost) could be potentially negotiated and managed on behalf of the entire scientific community.
2. A set of tools of wide and general utility that can be used by the applications. This would include a set of tools to manage data transfer, storage, and other data related activities, as well as coordinating a repository for useful software tools. The ideas outline by SciencePAD[2] could play a role here.
3. A set of tools to allow scientific communities to build a citizen-cyberscience facility where that is appropriate and useful.

Investment in Software

In addition to the above, an organisation put in place to coordinate such a set of coherent services and activities would also be the natural way to broker new collaborations. One area that will be of strategic importance in the coming years will be a significant investment in software capability that will be absolutely essential to obtain the best performance from current and future computer and storage architectures. Many sciences today benefit from commodity CPU and storage, and this is likely to change as the consumer market shifts from PC's to tablets and smartphones. This investment in software is essential to maintain European competitiveness in this area, and should include coordination of existing expertise to the benefit of diverse communities.

[2] http://www.sciencepad.org

There may also be traditional software and tools at the application layer that would benefit from a European-wide collaboration. Examples here may include a mechanism to obtain better licensing conditions, or collaborations to build specific application software of general benefit to a broad community.

Relationship with the HPC Community

The relationship with the supercomputing community (HPC applications) should also be re-defined. There are two aspects to consider. The first applies to the frontier-science challenges that need the most significant HPC resources. In this case the HPC facilities should be viewed as scientific instruments in their own right that produce science data for their application communities. Today such large-scale simulations produce huge volumes of data. Those application communities are then naturally users of an e-infrastructure on which to distribute and analyse their (supercomputer produced) data, and those applications would also be scientific stakeholders in a general scientific e-infrastructure.

There are other aspects of HPC facilities that are complementary to the cloud and High Throughput resources. Some applications that require modest levels of an HPC resource may well be deployable on suitably configured cloud or data intensive computing resources. There are also workflows that cross HPC and data intensive computing resources and would benefit from an integrated service environment.

The HPC facilities and their scientific communities would also benefit from the underlying technologies mentioned above (the networks, federated identities, policy work, etc.).

Building the data continuum

A data continuum, that is to say a system capable of navigating the data evolution by linking the different stages of the data lifecycle, from raw data to publication is necessary to accelerate the rate of scientific discovery and increase the impact of research on society. Elements of the data continuum exist and a range of projects, including openAIRE[3] where CERN provides the Invenio[4] software technology that supports this open access repository and many more around the world, have created repositories for initially, publications, and now extending to data that can give good examples of what can be achieved. But these remain independent projects and have not been integrated into the overall e-infrastructure landscape.

Providing Leadership

In order to build such a long term and broadly scoped e-infrastructure to benefit the entire European community, we must leverage the tremendous assets that have been built up during the last decade: in particular the knowledge and skills, as well as the working prototypes of each of the core services noted above.

[3] https://www.openaire.eu/
[4] http://invenio-software.org/

What we envisage is a continuum addressing the needs of education and speculative innovation through to growing and established entities. There must be therefore a range of infrastructure and services to satisfy the needs of such a broad range of maturity of activities. This continuum must cover the different axes of financial models (user-pays, provider-pays) as well as infrastructure (industry supplied and in-house). User representation in governance is paramount and we envisage a user activity for e-infrastructures as a well-defined activity.

The development of new and novel services and software must be publicly funded, but as these become production ready and commoditised they should move into the industrial service operation.

Open source and open standards are essential to ensure maximal adoption, and to allow new entrants to be able to leverage the innovation made with public funding.

The platform should also enable business innovation for new services and new users to create wealth and employment.

Funding models

Past experience has shown that as the number of communities and activities that could benefit from European e-infrastructure continues to grow and evolve, there is no "one size fits all" solution that is appropriate.

Within the lifecycle of a given activity, it may be appropriate to have supplier financed resources for a time and then some subsidised resources as the activity evolves. Fully paid resources by the activity may come at a later stage.

In terms of e-infrastructure it is important that activities can rely on the resources and services so it is critically important that timeframes for the e-infrastructures be long and the funding stable.

Where it is expected that users should pay for the services, those services must be relevant and attractive in order for that to happen. The added value of the proposed services must be made absolutely clear. In order to remain attractive to the user communities the offering must evolve and adapt to the changing needs. This evolution can begin with a set of managed federated services that are recognised as a common need, with new services being added as commonalities are explored and understood. The proposed platform must provide solutions in a timely and relevant way. This is another reason why the user community stakeholders must be directly involved in the overall governance.

A model for moving from innovative development (where a funded activity may develop a service) to industrial operation is essential to avoid unproductive competition between publicly funded activities and commercial offerings. Organisations such as CERN have specialist knowledge of large scale tendering and coordination of tendering

and brokering that would enable cost efficiencies to be gained through the application of scale.

Governance

It is important that a future European e-infrastructure be driven by the scientific stakeholders. Some key strategic research communities could be selected that would drive the frontiers of the technologies in several different but complementary aspects. This is covered in a separate document[5].

[5] David Foster, Bob Jones, "Science Strategy and Sustainable Solutions; A Collaboration on the Directions of e-Infrastructure for Science", CERN-OPEN-2013-017, http://cds.cern.ch/record/1545615/files/CERN-OPEN-2013-017.pdf