BDEC Workshop

Strategic collaboration between HPC and Cloud to address big data in global scientific challenges

Maryline Lengert (ESA), Bob Jones (CERN), David Foster (CERN), Steven Newhouse (EMBL-EBI)

1.1 Introduction

The last decade has seen a tremendous growth in e-infrastructure and related activity in a number of research communities as a result of funding by the European Commission in the 7th (and earlier) Framework Programmes and by corresponding national investments. Consequent to this investment in capacity, research communities are presented with several individually excellent, but independent – cross-layer initiatives which present researchers with sometimes inconsistent technical approaches and disjointed managerial structures to achieving a production quality infrastructure. It is being widely recognised that this fragmented landscape has increased the complexity and reduced the willingness of research communities in their adoption of these e-Infrastructure services. Recently, a vision has emerged that addresses this fragmentation by proposing an 'e-Infrastructure Commons'[1], an open environment where researchers can flexibly discover and chose the services and service providers from either the public and private sector that they feel will best meet their needs.

Until recently, accessing research services has been a relatively closed static system with researchers applying to single local, national or European compute and storage service providers, usually through review process to receive (if successful) an allocation of resources on the designated systems. The advent of publicly available commercial cloud services has provided an alternative approach for researchers and research communities. This approach has been further developed within the Helix Nebula initiative through the initial engagement of European Intergovernmental Research Organisations (EIROs), seeing it as a tool to perform generic data transformation processes. The Helix Nebula Science Cloud also brings unique data / knowledge / tools in a cross-domain market place catalysing science data to be seen in a different (un-known) context. Today science communities (earth, life, physics, etc.) want access and integration of many data sets regardless of location in order to address societal grand challenges.

1.2 Funding, sustainability and business opportunities

Today, the majority of existing public e-infrastructures are supported by national/regional funding agencies and provide services that are free at the point-of-use. The financial support provided by the funding agencies is normally based on a fee linked to the cost of setting-up and operating a service rather than its level of usage. By introducing a pay-per-usage scheme as part of the overall funding model for the allocation of a fraction of the resources, as has been demonstrated within Helix Nebula, the funding agencies will have the information to be able to measure the level of usage of a service and whether it justifies their investments. In addition, implementing the pay-per-usage model will give some of the financial control to the users and they will favour those services which offer better value-propositions. The result of these changes to the e-infrastructure business model will reduce the total cost of service provisioning (processes building on digital data) and consequently contribute to their sustainability. The move to a federated marketplace model was described within the 'Strategic Plan for a Scientific Cloud Computing Infrastructure for Europe'[2] in what became the Helix Nebula Initiative and was generalised in 'e-infrastructure for the 21st Century'[3] issued by the EIROforum IT Working Group.

[1] http://www.e-irg.eu/images/stories/dissemination/white-paper_2013.pdf

[2] http://cds.cern.ch/record/1374172/files/CERN-OPEN-2011-036.pdf

[3] http://dx.doi.org/10.5281/zenodo.7592

1.3 Vision for an e-Infrastructure Commons Marketplace

The e-commons infrastructure marketplace, driven by the European Research Area, will provide public and private researchers with access to worldwide and world-class resources and services through a dynamic and sustainable marketplace. This overarching infrastructure, built on public and commercial assets, will cover the entire scientific workflow from research to production, from problem-solving to discovery and innovation. The marketplace will offer the broadest range of services available today and will participate in the development of those needed for tomorrow. It will ensure use of open standard and interoperability of service providers while adhering to European policies, norms and requirements.

To achieve this vision requires:

- More coherence and integration from services providers (public and private) in the e-Infrastructure Commons marketplace
- To engage researchers in all disciplines from all sizes of community
- To keep resources free at the point of use for researchers
- To link resource use to service provider income for sustainability
- To reduce the barriers to entry and simplify use for end-users.
- A holistic view of pan-European existing and planned e-infrastructure.

The marketplace should encompass both publicly funded and commercial assets so that the sum of these e-infrastructures, with all their complementarity and variety of "circles of influences", will create a new momentum in Europe, driven by science, to implement a knowledge-based society and economy.

1.4 Expected impact

The expected impact of this Marketplace is:

- Researchers, supported by large scale long term research infrastructure, drive the evolution of services for their research needs
- Funding agencies benefit from market forces to establish volume and price
- Create a fertile environment that nurtures new scientific ideas and challenges
- Service providers are able to attract revenues to sustain services
- It establishes an ecosystem that benefits downstream industry
- It assembles an ever growing marketplace building on Information as a service based on federation of data and IP meeting European security and integrity requirements
- It provides visibility and incentives to industry to invest in new assets (as a business case but also to use the science communities for testing cutting-edge technology as has been demonstrated by the CERN openlab project[4])

A governance and operational model will integrate and unify these services and stimulate expansion and adoption to new research communities, new service providers and the integration of new innovative technologies. The governance model shall involve all the stakeholders, including service suppliers and service consumers (end-users), as well as funding bodies seeking to use this platform as a policy implementation tool, to ensure that the market remains open and competitive.

In this context, Cloud services can be seen as a research tool demonstrating how such models could be used within the HPC environment. Indeed the ability to access massive datasets (e.g. Climate change related data) will allow users to take advantage of the distributed aspect of the cloud and investigate parallel data-mining algorithms such as map-reduce techniques. It will also allow users with very limited bandwidth, such as users from developing countries, to access vast amounts of data, process them on the cloud and transfer much smaller amounts of summary information to their site.

[4] http://openlab.web.cern.ch/becoming-sponsor