Cloud Computing for Research & Innovation

David Fergusson, Francis Crick Institute
Martin Hamilton, Jisc (editor)
Philip Kershaw, STFC, CEDA & JASMIN
Steven Newhouse, EMBL-EBI & ELIXIR
Jacky Pallas, UCL, Farr Institute & eMedLab
Jeremy Yates, STFC DiRAC & SKA
## Contents

Executive Summary ........................................................................................................... 3  
Recommendations .............................................................................................................. 4  
1. The UK National E-Infrastructure ................................................................................. 6  
2. Cloud Computing within the UK National E-Infrastructure .......................................... 9  
3. Recommendations ........................................................................................................ 14  
4. Derived Actions ............................................................................................................ 17  
5. Roadmap: A 5 year vision for Cloud in UK Research. ..................................................... 20  
Acknowledgements .......................................................................................................... 23  
Annex A: What is the Cloud? A functional view ................................................................. 25  
Annex B: Cloud Computing for Researchers .................................................................... 30  
Annex C: Trust & public cloud ......................................................................................... 36  
Annex D: NeI and cloud .................................................................................................... 38
Executive Summary

Today, research across diverse domains such as physics, engineering, life-science, the environment and social sciences is being driven increasingly by the ability to collect, store and analyse large data sets – so called ‘big data’. The UK’s National E-Infrastructure (NeI) needs to integrate through high-bandwidth low-latency networks the computational, data and storage services needed by researchers to support their ‘big data’ analysis to rapidly carry out their world-leading collaborative research programmes.

A key component of the NeI is cloud computing – the elastic, on-demand provisioning of infrastructure, platforms or software – to meet the needs of researchers from both the public and private sectors. Such a hybrid model requires integration of public sector institutional, community and national resources with those available internationally in both the public and private sector.

Given the strategic importance of the NeI, and the growing importance of cloud computing for big data analytics in the research community, members of the e-Infrastructure Project Directors Group (PDG), at the request of the RCUK National e-Infrastructure Group were asked to identify a set of technical and policy recommendations that will improve the accessibility and usability of cloud resources - for research, teaching and administration.

This report identifies the major technical and policy issues that are seen to be preventing widespread take up of cloud services for the UK academic and related community and provides a 5 year roadmap to investigate these issues and provide closer integration of public and private sector resources to improve the capability of the UK research community.

David Fergusson, Francis Crick Institute
Jacky Pallas, UCL & eMedLab
Martin Hamilton, Jisc (editor)
Philip Kershaw, STFC, CEDA & JASMIN
Steven Newhouse, EMBL-EBI & ELIXIR
Jeremy Yates, STFC DiRAC & SKA

(We are indebted to a number of colleagues for their contributions to this report – please see the Acknowledgements section for a full list of acknowledgements.)
Recommendations

1. Provide a clear community focus for cloud computing within the UK NeI by establishing a Cloud Computing Working Group that reports directly to the RCUK NeI Group.

A Cloud-Special Interest Group should also be set up to provide a forum for all members of the UK research community (both consumers and providers) to come together to exchange best practice and experiences. The SIG would be able to provide a grass roots view from across all research domains as to future cloud computing needs, which can then be taken up and turned into a strategy by the Working Group and communicated to RCUK NeI Group. The Cloud Computing Working Group would be able to establish smaller focused working groups to discuss specific issues as required.

An important task of this group would be to exert influence on the providers of cloud software and scientific software vendors to include the requirements of the UK research community in future releases.

2. Provide a minimal technical integration of the UK NeI resources to promote workload mobility and to reduce technical barriers to entry.

These technical barriers can be reduced by establishing a consistent access model across all UK NeI resources. These should be implemented so as to enable researchers to use a single set of identity credentials to access services at their home institution and to access and move their workloads between local, regional, national and international cloud computing resources as they require. An integrated authentication and authorisation infrastructure (AAI) is needed alongside consistent open interfaces to the resources so as to make workloads mobile between different cloud computing providers. This will foster competition and prevent lock-in.

Considerable investment has already been made in federated AAI initiatives such as Moonshot (which is now available as the Jisc Assent service), but further investment is needed to integrate these AAI technologies into cloud computing platforms and to make NeI resources (and those in the private sector) available.

Virtual research environments have an important role to play enabling researchers to create software environments tailored to the needs of their application domain and user communities. At the same time however, there is a need to facilitate the sharing of application environments to enable workloads to be easily deployed on any UK NeI compliant cloud. Container technologies such as Docker are an important enabler for the development of this capability.

3. Equipping the research community with the right skills and support to fully exploit UK NeI cloud resources.

Many NeI service providers are finding it challenging to find staff with the right skills to operate their cloud infrastructures and to provide the consultancy necessary for research groups to rapidly and successfully exploit
the available cloud resources. RCUK needs to invest in both basic and advanced training for service providers and those working directly with researchers to support their move to cloud computing resources. Investing in staff working in these fields provides them with skills that are potentially very transferrable into private sector as part of normal staff migration. In particular, training needs to be given to those working to support researchers in accessing cloud infrastructures – most likely the IT Support groups at an institution. Recent experience with the deployment of private and community clouds in the UK research community has highlighted the need for the devops role, a person whose skill set bridges the traditional system administration and software developer roles.

4. **Policy changes needed within RCUK to grow the adoption of cloud computing and the policy actions that RCUK can initiate externally on behalf of the UK cloud computing community.**

Cloud computing is acquired on an on-demand basis as required (operating expense) as opposed to an initial upfront payment (capital expense). RCUK funding models need to adapt to reflect this change for both researchers and community service providers who may consume commercial cloud resources in a hybrid model. Outside of the public sector, RCUK must continue and develop the activities that have been initiated through Jisc in establishing terms and conditions with commercial cloud providers to explore how the buying power of the community as a group can make purchasing of these services more effective, efficient and productive.
1. The UK National E-Infrastructure

1.1 Context

The context for the National E-Infrastructure is laid out in *A Strategic Vision for UK e-Infrastructure a roadmap for the development and use of advanced computing, data and networks* (Tildesley, 2012)

The main recommendations from the Tildesley report are summarised as follows:

1. Create a ten-year roadmap to define the components of the infrastructure: networks; data and storage; compute; software and algorithms; security and authentication; people and skills.
2. Create secure data and information stores in strategic locations with data-analysis provided through cloud environments, working with open source software.
3. Ensure that important public databases are available to all UK researchers
4. Provide broad access to the infrastructure for industrial partners, suppliers and Independent Software Vendors (ISVs), as well as the academic community.
5. Assist the development of a portfolio of training modules in computational science, numerical algorithms grid-computing, parallel programming, cloud computing, data-centric computing, e-science, computer animation and computer graphics.
6. Develop a single coordinating body to drive closer cooperation and enable effective industrial access, while insuring that UK academe has access to leading edge capability.

Investments by BIS, the Research Councils and HEIs in 2011-12 (£160M), 2012-2013 (£189M) and 2014-15 (£257M) have resulted in core elements of this vision (shown above in bold) being put in place.

1.2 Definition

The National E-Infrastructure (NeI) are those resources, linked by high bandwidth networks, which provide UK researchers with the computational, data and storage services they need to carry out their world-leading collaborative research programmes. In the future, these resources, which may be located in both the public and private sector, need to be integrated to allow researchers to access much greater capacity and capability than they possibly locate in their own institutions.

An integrated NeI will allow researchers from all disciplines to:

---


• Expand the complexity of their simulations/analysis to produce better science.
• Decrease the time needed to obtain these science results.
• Provide a technology platform that will support innovation in both the public and private sector.

The RCUK National e-Infrastructure Group (Chair, Morrell), comprised of the 7 research councils, IUK, Jisc and the Met Office, provides strategic oversight of the activities of the NeI Projects through the NeI Project Director’s Group. The NeI Project Directors Group (Chair, Yates), which comprises representatives from NeI Projects and Providers (see Annex D) and representatives of Innovate UK, Jisc and the HPC-SIG, is the body charged with integrating the projects and providers in such a way that an authorised researcher can access her/his resources via a simple interface and using only one set of credentials.

There are over 20 Large and Specialist projects that have been set up within the NeI ecosystem including the Janet6 high bandwidth low latency network, data and compute services for social and economic science, genomics in life sciences and medicine, climate research, particle physics and innovation, and compute services in particle physics. In addition there are over 35 HEIs providing services to their resident academics.

The current state of the National E-Infrastructure is described in The National E-Infrastructure 2014 Survey3 (Yates & Hamilton, NeI Project Directors Group, 2014) and its future evolution is described in: The E-infrastructure Roadmap4 (Morrell, Chair, RCUK NeI Group, 2014).

1.3 Requirements

In order to establish a researcher centric infrastructure it is necessary for:

• The researcher to discover and access to a broad suite of integrated compute, data and storage resources that are accessed by a high-bandwidth low-latency network.
• To develop virtual research environments to more readily enable researchers to transition from desktop application experience to services deployed on remote infrastructures. This is essential for the long tail of science research to effectively exploit big data.
• A virtual research environment that allows researchers to discover the available and accessible resources, to move data between resources, to run reproducible and publishable workflows that supports open science and open data.
• A common UK Authentication Infrastructure that is interoperable with international identity management infrastructures, so allowing the user to use NeI resources using a single set of identity credentials.
• An Authorisation and Allocation/Accounting Infrastructure that allows research domains and projects to authorise researchers to use appropriate resources, allocate those resources and measure their usage.

3 http://hpc-sig.org/?wpdmdl=492
4 https://www.epsrc.ac.uk/newsevents/pubs/e-infrastructure-roadmap/
• A secure network and storage environment that can offer at rest and in flight information assurance to those research projects/communities that have data security concerns.
2. Cloud Computing within the UK National E-Infrastructure

2.1 Importance of Cloud Computing for Researchers

Cloud computing has enormous potential as an enabling technology for the research community. Computing resources can be provisioned on demand on as needed basis and can be made elastic to grow and shrink as a given workload requires. Perhaps most important, it provides a key solution to the challenge of big data by bringing users to the data. In this model, cloud enables the provision of virtual computing environments co-located with centres capable of hosting the huge amounts of data associated with many research activities. These can provide vastly more computing capability than available in users’ home institutions and avoid the need for the transfer of large data volumes.

Activities are categorised by deployment model (private, public, community or hybrid clouds) and by the service they offer (infrastructure, platform or software) to the user – See Annex A for more details. Many commercial cloud computing services that are now offered to ordinary consumers can benefit members of the research community – See Annex B for more details.

2.2 Current Status

The Cloud Computing Working Group was established in August 2013 following an action from the National e-Infrastructure Project Directors Group. Its role is to foster collaboration and establish best practice for the application of Cloud Computing in the UK research community. The executive summary provided by the working group (Sept 2013) identified a number of areas that need to be addressed:

- There is no obvious entry point for research users, co-ordinated and training and guidance is needed
- A number of centres have plans to (or are in the process of) rolling out Private Cloud infrastructures. There is a need to co-ordinate and share experience between groups to establish best practice and avoid fragmentation and duplication
- There is a need to engage and work with commercial Public Cloud providers and co-ordinate usage with partners in the research community – in terms of technology and policy.

A workshop was organised in November 2013 to bring together members of the research community and Cloud computing experts and practitioners to elicit feedback on the findings of the executive summary. The workshop

5 https://indico.cern.ch/event/281517/timetable/#all
showed that there is a strong latent interest in the research community to be better informed and guided on how best to make practical use of the technology.

Presentations were co-ordinated around four key themes: use of Public Cloud, deployment of Private Cloud by e-infrastructure facilities, cloud federation and brokering – the bridging together of resources from different Cloud providers.

Table 1 lists the main cloud projects funded by UK or EU agencies that are relevant to researchers in the UK. With the increasing cross-border nature of our research collaborations and research projects the UK has an opportunity to play a leading role in such initiatives and our own domestic clouds must interoperate with these larger cloud infrastructures.

Table 1: A sample of UK and EU funded Cloud Projects

<table>
<thead>
<tr>
<th>Project</th>
<th>Cloud Technology</th>
<th>Main Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>EGI (EU)</td>
<td>Various</td>
<td>Provide a single federated cloud service from many individual infrastructure providers from across Europe.</td>
</tr>
<tr>
<td>EMBL-EBI – Embassy Cloud</td>
<td>VMware</td>
<td>Data Analysis for Bioinformatics.</td>
</tr>
<tr>
<td>EUDAT (EU)</td>
<td>OwnCloud</td>
<td>Provide a standard set of services for movement and storage of data built on top of cloud infrastructure</td>
</tr>
<tr>
<td>World LHC Compute GRID (STFC)</td>
<td>Openstack (CMS)</td>
<td>Computational and Data Services for LHC</td>
</tr>
<tr>
<td>Square Kilometre Array (STFC)</td>
<td>Openstack</td>
<td>Computational and Data Services for the SKA</td>
</tr>
<tr>
<td>Helix Nebula (EU)</td>
<td>Various Commercial and Open Source</td>
<td>Create a federation of providers and a market place for European scientific application domain</td>
</tr>
<tr>
<td>JASMIN2 (NERC)</td>
<td>VMware</td>
<td>Data analysis environment for the environmental sciences community</td>
</tr>
<tr>
<td>CLIMB (MRC)</td>
<td>Openstack</td>
<td>Computational and Data Services for Microbial DNA analysis</td>
</tr>
<tr>
<td>eMedLab (MRC)</td>
<td>Openstack</td>
<td>Computational and Data Services for Human DNA and disease analysis</td>
</tr>
<tr>
<td>Cambridge-AWS Link</td>
<td>AWS</td>
<td>Test for Hybrid Cloud job submission</td>
</tr>
<tr>
<td>NECTAR (Australia)</td>
<td>Openstack</td>
<td>Australian government funded cloud to support Australian research community6</td>
</tr>
<tr>
<td>EU To</td>
<td>Unknown</td>
<td>Proposal to create a hub of knowledge and expertise to coordinate technological developments to meet the</td>
</tr>
</tbody>
</table>

---

2.3 Next Steps

The recent PDG report *Imaging the UK National Data Infrastructure*\(^7\) suggested that the use of Cloud Technologies is a pre-requisite for the creation of a coherent National Data Infrastructure.

The Self-Service nature of Cloud and the ability to orchestrate resources make this technology particularly relevant to data driven science.

Cloud is widely seen as the **next-generation** IT delivery model

- Agile & Flexible
- Utility-based on-demand consumption
- Self-service driving down administrative overhead and maintenance

Public clouds are setting the benchmark of how IT could be delivered to users

- **However not all organisations and/or workflows are ready for public cloud**

Applications are being written differently today

- More tolerant of failure
- Making use of scale-out architecture

Our data is too large

- Volumes of data are being generated at unprecedented levels
- Most of this data is unstructured

Service requests are too large

- The time to science could get much longer without access to elastic resources
- More and more devices are coming online

---

7 [http://www.eu-t0.eu/](http://www.eu-t0.eu/)
• Tablets, phones, laptops, BYOD generation...

Crucially, applications weren't written to cope with the demand!

• Traditional infrastructure capabilities are being exhausted
• Service uptime, QoS, KPI’s and SLA’s are slipping

2.4 The Current Cloud Picture in the NEL

The fragmented and silo-ed nature of the NEL at first sight suggests that the resources could be more effectively used using cloud technologies.

In theory workloads could be distributed among the many systems of the NEL.

However several problems present themselves:

• The assets of the NEL are owned by particular research domains and projects.
  o There is no incentive to share resources and no incentive to consolidate resources.
• There is no common identity management system.
• The spare capacity is simply not there in the NEL.
  o Elasticity, which should be a benefit of cloud, is not present in a very full NEL.
• Cloud is a new technology and many of our systems staff and users are unfamiliar with it.
• There is no financial model of the NEL.
  o There is no resource brokering service.
• There is already community that is very hard to size that makes use of public cloud provision.
  o This suggests the NEL is not working for these people.

Progress, however, has been made:

• AAAI has been invested in with significant buy-in from across the Research Councils
• Jisc have created relationships with public cloud providers, such as AWS and Google, to create management portals to these services.
• Jisc have partnered with Microsoft⁹ and AWS¹⁰ to directly peer the Janet network with their datacentres, facilitating the adoption of Hybrid Clouds that mix institutional and cloud provider resources.
• The Investments by BIS and the Research Councils have created a sense of coherence and community that did not exist 3 years ago.
  o The creation of Private Clouds in the areas of the Environmental sciences, LHC and SKA data processing, and Bioinformatics.

---


¹⁰ https://www.jisc.ac.uk/amazon-web-services
o The creation of Software as a Service (SaaS) infrastructures such as the NERC Environmental Workbench, the Environmental genomics – Virtual Desktop.

o DiRAC and The National Service Archer, already share infrastructures and represent the Physical Sciences and Engineering groupings in the UK.

o DiRAC and GridPP have already swapped projects between them manually in order to make best use of resources.

o The co-operation of the Social Scientists, Medical Bioinformatics and Patient Records projects in working with Jisc to produce Safe Share is an example of how difficult projects such as Cloud uptake can be managed in the future.

o The Infinity Data Centre in Slough has created an environment in which co-location of equipment is now possible and desirable.

o The leading role of the UK in GÉANT and the EGI should make sure that the UK builds structures that interoperate with and leverage resources made available via EU projects.

o The improvement to Janet Network speed and latency, and security, means that systems (and data) can now be distributed. This has been shown to work by GridPP who can now use the Network as “mobile” storage and regularly are moving PBs a week over Janet.

o The possible future ability of the Janet Network itself to operate in Self-Service mode is a welcome addition to the Cloud deployment model.

- Jisc and the PDG have surveyed the attitude of the Nel and HEI CIOs to Cloud technologies\textsuperscript{11}
- Courses are becoming available to introduce people to the Cloud as an operating system, including ones specifically targeting researchers, such as those run by Microsoft in partnership with the Software Sustainability Institute and Oxford e-Research Centre\textsuperscript{12}.


\textsuperscript{12} http://research.microsoft.com/en-us/projects/azure/training.aspx
3. Recommendations

Although a number of needs have been identified, the members of the Cloud WG have limited resources to address these. A co-ordination role has been proposed to assist the work of the working group chair and committee to organize and facilitate meetings.

3.1 Community Building

- Set up a RCUK NeI Cloud working group and SIG – governance arrangements to be specified. The working group is responsible for the delivery of the (fully resourced) roadmap.
- Liaise with groups in the research community who are rolling out Private Cloud interfaces to their e-infrastructure. Share experience and establish best practice in the use of software frameworks, resourcing for deployment and operations of services.
- Build the community in the UK – present a single face to funds, stakeholders and vendors.
- The working group should identify pilots to be supported – RCUK budget requested for cloud innovation fund.
- Share the knowledge and experience of successful projects in different subject domains to ensure transferability.
- Document examples of the research community’s experiences of using public cloud providing guidance on what works, and what doesn’t by seeding activity by surveying researchers (recent RCUK grant award PIs?) on their use of public cloud.

3.2 Technical Integration

- If it is to succeed, any technical integration must take its lead from science-driven use cases. Clear focus is needed on which interfaces are required and who is the target user, be they an administrator, software developer or end user researcher.
- Co-ordinate with activities underway to broker or federate resources from multiple Cloud providers and assist e-Infrastructure providers and end-users to best exploit these initiatives. These activities include JANET Cloud, G-Cloud, EGI Federated Cloud Task Force and Helix Nebula.
- Target small amounts of funding to develop UK priorities that aren’t being met elsewhere, seed and influence product evolution.
- Help the UK to make the best use of community and public clouds – e.g. portability of applications and workloads. The purpose of this is to ensure that scientific workloads can be executed on the most efficient and effective infrastructure. The goal is an ecosystem of clouds across domains and providers, supporting inter-disciplinary science.
3. Recommendations

- Project to explore common cloud management portal – link to AAAI.
- Project to explore emerging container technologies such as Docker as standard/recommended approach to packaging and portability. We need to identify and prioritise use cases for application of this technology.
- Rather than a single centralized research cloud, but instead looking for opportunities to collaborate from efforts already underway in the individual research domains making more effective use of UK infrastructure and public cloud, such as eMedLab and CLIMB, and sharing/collaborating e.g. around documentation.
- Cost and timing aspects of data storage and egress – e.g. 500TB from AWS would cost roughly $150,000 per annum to store, and $25,000 in egress charges to export from the cloud. Large scale data transfer required to move workloads from e.g. EBI to other facilities requires specialist network infrastructure.
- Access and sharing regime - limitations in Ceph and Swift, as focus has been on high performance. Controls around sharing images are also somewhat absent. How do we act collectively to change this? It will happen eventually by itself, albeit perhaps more slowly than we would like.
- We need to develop the concept of virtual research environments: web-hosted applications that offer easy to use interfaces that can be tailored and customised to meet the needs of users who are more used to point and click desktop applications.
- We should also offer to our users the option of virtual desktops. These are simply virtual machines that are generated and allow the users to run their desktop/laptop applications and workflows on much more powerful systems.
- This would help users to transition from desktop applications that they are more familiar with to cloud hosted applications. This class of user represent our main access problem and they actually would come from across the research spectrum.
- There will always be those users who want to login a system and submit jobs from the command line. That should not be discouraged.

3.3 Training and Support

- Training needs to be carefully targeted to the right class of users. End user applications are likely to remain within the scope of a given application domain. However, at the platform and infrastructure tiers there is a need to train and equip administrators and a new class of devops users.
- Support end users in the research community to assist them in how to use services from Cloud providers and co-ordinate with training initiatives through the PDG members.
- For end users, provide a coherent route from desktop application environment through to virtual environments hosted on e-Infrastructure, including community and public cloud.
- A central programme of support akin to ARCHER CSE including training, advice and guidance for exploiting public cloud, working with cloud providers.

• The Research Software Engineering (RSE) community\(^{14}\) has a pivotal role to play in helping researchers to exploit the potential of cloud technologies for research and innovation - for example by porting common packages and environments to run in the cloud, and packaging with Docker or other suitable container technologies.

3.4 Policy Issues

• Liaise with Public cloud providers on behalf of e-infrastructure facilities and end-users in the areas of: training – how can users’ best use public cloud for research, best use of technology to fit with research workloads, policy (SLAs) and funding.
• Further cloud brokerage/peering work by Jisc – building on AWS model with other providers
• Common institutional level management / billing portal, discounts for bulk usage, delegation of budget to PIs, self-service management of researchers by PI / administrator, tools for tracking spending
• Cost characterisation – refreshing the work of the Jisc/EPSRC study from 2011\(^{15}\)
• Common approach for use of public cloud in grant proposals (extends to wider NeI?)
• Establish “public cloud as an NeI resource” (Jisc could broker access to this?)
• Policy statement on what data can be put into cloud providers, compliance requirements

\(^{14}\) http://www.rse.ac.uk/
\(^{15}\) https://www.epsrc.ac.uk/research/ourportfolio/themes/researchinfrastructure/subthemes/einfrastructure/cloud/
4. Derived Actions

Tables 2 and 3 are a list of suggested actions

**Table 2: Actions for the Future from the Infrastructure perspective.**

<table>
<thead>
<tr>
<th>Action</th>
<th>Responsible Party</th>
</tr>
</thead>
<tbody>
<tr>
<td>Produce a financial model of the NeI. This is the first step for any resource brokering model.</td>
<td>RCUK</td>
</tr>
<tr>
<td>Produce a logical map and description of the NeI and its services</td>
<td>Jisc and PDG</td>
</tr>
<tr>
<td>Agree AAAI roll out across PDG</td>
<td>PDG, Jisc, and RCUK</td>
</tr>
<tr>
<td>1. Adoption of appropriate Authentication model</td>
<td></td>
</tr>
<tr>
<td>2. Adoption of Authorisation and Accounting Model</td>
<td></td>
</tr>
<tr>
<td>3. Need to Federate with external Access Management Provider</td>
<td></td>
</tr>
<tr>
<td>Ask PDG members to suggest possible private clouds and suggest how investments could be better co-ordinated.</td>
<td>PDG</td>
</tr>
<tr>
<td>What can we do to consolidate and build economies of scale to help make ourselves much more elastic?</td>
<td></td>
</tr>
<tr>
<td>Network as a Service models need to be understood and deployed</td>
<td>Jisc</td>
</tr>
<tr>
<td>Assess Cloud Brokerage Models</td>
<td>Jisc</td>
</tr>
<tr>
<td>Deployment of Network Security services</td>
<td>Jisc</td>
</tr>
<tr>
<td>Understand how the NeI can be made elastic by resource pooling and use of Public Cloud</td>
<td>PDG</td>
</tr>
<tr>
<td>Ask RCUK to set up a Cloud Working Group to inform and guide the adoption of Cloud in the NeI.</td>
<td>RCUK</td>
</tr>
</tbody>
</table>
4. Derived Actions

**Produce a financial model of the NeI. This is the first step for any resource brokering model.**

<table>
<thead>
<tr>
<th>Action</th>
<th>Responsible Party</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complete NeI to Public Cloud tests at Cambridge. This will provide a prototype for the NeI</td>
<td>Cambridge and AWS</td>
</tr>
<tr>
<td>Have a meeting on Cloud Technologies and get consensus on the projects and Cloud Technologies the UK should be supporting.</td>
<td>PDG</td>
</tr>
</tbody>
</table>

These Projects will include:

1. Suggestions for sensible deployment stacks
2. Which technologies should we support
3. NeI interfaces with Cloud Technology – Self Service, and metering
4. AAAI interfaces with Cloud Technology
5. Understanding which areas will offer SaaS, PaaS and IaaS
6. A roadmap for using Cloud for HPC
7. How can we develop Cloud technologies, particularly in the areas of data management
8. Can we make Hybridisation easy

**Publish Reports on NeI and the Cloud and submit to RCUK and ELC**

<table>
<thead>
<tr>
<th>Table 3: Requirements from a researcher point of view</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Researcher requirement</strong></td>
</tr>
<tr>
<td>---------------------------</td>
</tr>
<tr>
<td>Single Sign on</td>
</tr>
<tr>
<td>Researcher requirement</td>
</tr>
<tr>
<td>---------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Knowing where Resources are</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Accessing those resources</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Collaborative analysis and processing environments, ability to share data, analysis and</td>
</tr>
<tr>
<td>results with their peers</td>
</tr>
<tr>
<td>Building Workflows</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Producing Virtual images and Instances</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Running My work flow</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>
5. Roadmap: A 5 year vision for Cloud in UK Research.

Cloud will,

1. Aid the UK economy by allowing UK researchers to do more with less and provide a competitive data infrastructure underpinning UK research, cutting across research sectors and allowing industry to access the same e-infrastructure and its resources. This should increase the productivity of the UK research sector and UK Industry.

2. Enable the sharing of research and knowledge within the UK to improve efficiency and effectiveness, with impact that benefits the UK economy.

3. Provide World Class infrastructure for UK research. Easily accessible cloud infrastructures for everybody.

4. Meeting the challenges of data intensive research and the development of data science, particularly the Alan Turing Institute.

This will be done by following a roadmap that will allow our e-infrastructure to make optimal use of cloud technologies to deliver our research services. By understanding which compute workloads (e.g. High Performance Computing, High Throughput Computing, database analysis) could move to the public cloud, and making "private cloud" facilities cloud compatible with the public cloud, we can allow the end user to define their own routes to IT services through to achieve their compute goals and pick the appropriate service provider(s).

Roadmap

Year 1
<table>
<thead>
<tr>
<th>Task</th>
<th>Actions</th>
<th>Stakeholders and Resource</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Movement</td>
<td>Replicate large scale data from data sources to sinks in a secure fashion, e.g. EBI to eMedLab, EBI to CLIMB</td>
<td>Jisc, eMedLab, CLIMB, EBI. 1FTE for 3 months</td>
</tr>
<tr>
<td>Barriers to using Public Cloud</td>
<td>Explore issues around use of public cloud, e.g. egress charging</td>
<td>Jisc, WG, 1 FTE for 3 months</td>
</tr>
<tr>
<td>Brokerage</td>
<td>Brokered community agreement with public cloud providers. This requires clear consistent/unified statement from RCUK about how “capital” funds can be used to purchase public cloud capacity.</td>
<td>Jisc, WG, RCUK, 1 FTE for 3 months</td>
</tr>
<tr>
<td>Track Existing Activities</td>
<td>UK NeI Cloud projects, EU Projects, Public Cloud usage</td>
<td>Jisc. WG</td>
</tr>
</tbody>
</table>

**Year 2**

<table>
<thead>
<tr>
<th>Task</th>
<th>Actions</th>
<th>Stakeholders and Resource</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brokerage Service</td>
<td>Brokered cloud facilities start to become available to researchers</td>
<td>Jisc/0.3 FTE p.a.</td>
</tr>
<tr>
<td>Group based access to resources with single identity</td>
<td>AAAI implementation available to NeI Projects and a API is developed to allow use by Public Cloud Providers’ AAAI infrastructure . Work is already in progress – Safe Share and the integration of Assent and SAFE. A pilot has already allowed ABFAB (Assent) to interface with Openstack Keystone.</td>
<td>Jisc, WG, EPCC, ADRC, Farr, eMedLab. 4FTE per annum in 2015-2016</td>
</tr>
</tbody>
</table>

**Year 3**
## Task Actions Stakeholders and Resource

### Standard Services Available

<table>
<thead>
<tr>
<th>Actions</th>
<th>Stakeholders and Resource</th>
</tr>
</thead>
<tbody>
<tr>
<td>General On demand Self Service available an option for NeI users for Compute and data services</td>
<td>WG, Jisc, RCUK, Research Domains (2-3 FTE for 2015-2106)</td>
</tr>
<tr>
<td>Management API</td>
<td>WG, Jisc (1 FTE for 1 year)</td>
</tr>
<tr>
<td>Standardised workflow images (or containers)</td>
<td>Research Domains (Resource TBD)</td>
</tr>
<tr>
<td>Secure data access within private cloud, including improvements to container security</td>
<td>WG, Jisc (2.5 FTE p.a.)</td>
</tr>
</tbody>
</table>

### Year 4

<table>
<thead>
<tr>
<th>Task</th>
<th>Actions</th>
<th>Stakeholders and Resource</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cloud Federation</td>
<td>Federated data access between private clouds</td>
<td>Jisc, WG, RCUK 2 FTE for 2016-17</td>
</tr>
</tbody>
</table>

### Year 5

<table>
<thead>
<tr>
<th>Task</th>
<th>Actions</th>
<th>Stakeholders and Resource</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hybrid Model Available</td>
<td>Hybrid model that lets users exploit the best of community specific private cloud and public cloud resources, and seamless/efficient migration of workload and data between federated clouds. Expectation that containers will play a major role in this</td>
<td>Jisc, WG, RCUK 2 FTE for 2017-18</td>
</tr>
</tbody>
</table>
Acknowledgements

We would like to thank the following people for giving up their time to provide input into this report:

Neelofer Banglawala, EPCC
Stephen Booth, EPCC
Brendan Bouffler, Amazon Web Services
Craig Box, Google
Peter Boyle, University of Edinburgh / DiRAC
Peter Braam, University of Cambridge / SKA
Steven Bryen, Amazon Web Services
David Britton, University of Glasgow / GridPP
David Chadwick, University of Kent
Tim Chown, Jisc
Jeremy Coles, CERN
Ian Collier, STFC
David Colling, Imperial College
Tim Cutts, Wellcome Trust Sanger Institute
Shaun de Witt, STFC
Ruediger Dorn, Microsoft
Matthew Dovey, Jisc
Tom Eddington, National Oceanography Centre
William Florance, Google
Paul Fretter, Norwich BioScience Institutes
Andy Grant, Atos
Scott Hamilton, Amazon Web Services
James Hetherington, UCL / Software Sustainability Institute
Terry Hewitt, STFC Hartree Centre
Adam Huffman, Francis Crick Institute
Laurence Hurst, University of Birmingham
Ed Jackson, Amazon Web Services,
Claire Jenner, UCL
Jens Jensen, STFC
Owain Kenway, UCL
Iain Larmour, EPSRC
George Leaver, University of Manchester
Peter Maccallum, Cancer Research UK Cambridge Institute
Simon McIntosh-Smith, University of Bristol
Matt McNeill, Google
Aleks Nenadic, University of Manchester / Taverna
Dan Perry, Jisc
Robin Pinning, University of Manchester / N8
Alan Real, University of Leeds / HPC-SIG / N8
Barak Regev, Google
Thom Rödde, Google
Mark Rowlands, Amazon Web Services,
Andy Richards, University of Oxford / OERC / SES5
David Salmon, Jisc
Jeremy Sharp, Jisc
Rhys Smith, Jisc
Kenji Takeda, Microsoft
John Taylor, University of Cambridge
Andy Turner, EPCC
Simon Thompson, University of Birmingham / CLIMB
Ash Vadgama, AWE
Annex A: What is the Cloud? A functional view

There is widespread misunderstanding about the term ‘Cloud computing’ leading at times to accusation of hype or at other times conflation with established computing paradigms such as Grid Computing. An understanding of the concepts is essential in order to make informed choices about how it can be best exploited and in what ways it differs or complements what has gone before.

Cloud computing powers the services of Internet giants like Microsoft, Google and Amazon. That technology is now available to institutions, to learners and to researchers. This is hugely empowering, for example by extending the reach of an individual far beyond what would historically have been possible - for a few tens of dollars a junior researcher, undergraduate or citizen scientist can create the equivalent of a million pound supercomputer for a few hours to carry out a calculation.

Cloud computing is a portmanteau term encompassing everything from infrastructure as a service (essentially renting someone else’s server equipment) through to software as a service (typically websites that someone else runs for you). In the middle, there is a platform tier that provides the micro-services that power the likes of Android and iPhone apps, and also many web delivered services.

But it is also important to take an informed view of cloud services, particularly where risk management, switching costs and sustainability are concerned. However, at the same time there is a significant amount of fear, uncertainty and doubt about cloud computing - and some genuine concerns.

Today’s concept of cloud computing grew out of two parallel activities that eventually converged - commercial hosting of servers and storage in professionally run data centres, and the two key realisations by leading Internet firms. These were that a) they could rent out capacity that would otherwise be spare, and b) opening up their application programming interfaces (APIs) to third party developers would make it possible to create a vibrant ecosystem of applications that no one company could likely construct on its own. It is also commonplace for mobile apps to be built up on a substrate of cloud services, even if this is not readily apparent to the end user.

In the experience of authors of this report, the US National Institute of Standard and Technology (NIST) Definition of Cloud Computing16 document provides an excellent starting point to understand this technology. For its definition it sets out three key areas:

- Essential Characteristics: what are the properties that allow us to say one system is a cloud where another is not? This is key to understanding capabilities of a given system and what are the differentiating factors that make cloud technology unique.

---

• Service Models: what services does a cloud offer and to whom? The three models described can be considered as successive abstraction above underlying computer hardware upon which a cloud runs. At the lower levels, cloud computing interfaces provide administrators with powerful capability to deploy whole virtual computing infrastructures. At the opposite end of the scale end users using an application may have no or little concept that it is cloud-hosted.

• Deployment Models: too many ‘cloud’ may be synonymous with large public cloud providers but a cloud may be deployed in any of number of deployment models including for example a private cloud hosted for the use of a single organisation. The deployment model or combination of deployment models are a critical consideration for getting the most from cloud computing technology for any given research project or programme.

In the following all text quoted in italics is taken verbatim from the NIST document.

Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction.

This cloud model is composed of five essential characteristics, three service models, and four deployment models.

**A.1 Essential Characteristics**

**On-demand self-service:** A consumer can unilaterally provision computing capabilities, such as server time and network storage, as needed automatically without requiring human interaction with each service provider.

**Broad network access:** Capabilities are available over the network and accessed through standard mechanisms that promote use by heterogeneous thin or thick client platforms (e.g., mobile phones, tablets, laptops, and workstations).

**Resource pooling:** The provider’s computing resources are pooled to serve multiple consumers using a multi-tenant model, with different physical and virtual resources dynamically assigned and reassigned according to consumer demand. There is a sense of location independence in that the customer generally has no control or knowledge over the exact location of the provided resources but may be able to specify location at a higher level of abstraction (e.g., country, state, or data centre). Examples of resources include storage, processing, memory, and network bandwidth.

**Rapid elasticity:** Capabilities can be elastically provisioned and released, in some cases automatically, to scale rapidly outward and inward commensurate with demand. To the consumer, the capabilities available for provisioning often appear to be unlimited and can be appropriated in any quantity at any time.

**Measured service:** Cloud systems automatically control and optimize resource use by leveraging a metering capability at some level of abstraction appropriate to the type of service (e.g., storage, processing, bandwidth, and active user accounts). Typically this is done on a pay-per-user charge-per-use basis. Resource usage can be
monitored, controlled, and reported, providing transparency for both the provider and consumer of the utilized service.

A.2 Service Models

Software as a Service (SaaS): The capability provided to the consumer is to use the provider’s applications running on a cloud infrastructure. The applications are accessible from various client devices through either a thin client interface, such as a web browser (e.g., web-based email), or a program interface. The consumer does not manage or control the underlying cloud infrastructure including network, servers, operating systems, storage, or even individual application capabilities, with the possible exception of limited user-specific application configuration settings.

SaaS is exemplified in the commercial sector by the likes of the Microsoft Office365\textsuperscript{17} and Google Apps for Education\textsuperscript{18} communications and collaboration suites, and the SalesForce.com\textsuperscript{19} Customer Relationship Management (CRM) system. Software vendors are increasingly moving to delivering applications through cloud computing as this reduces the friction of taking up the product - and more cynically helps them to retain customers, exploit customers’ data, and to upsell customers to other products and services.

In the research sector, there is a long history of web-based applications and services to support distributed user communities. These could be classified under the category of SaaS. In recent years, a number of initiatives have grown around the concept of Virtual Research Laboratories, the idea of shared workspaces for scientific user to collaborate and hosted on a cloud infrastructure to enable access to greater computing processing capacity and storage that would otherwise be possible with a desktop application. Examples are the CSIRO Virtual Laboratories hosted on the Australian research cloud, Nectar\textsuperscript{20}.

\textsuperscript{17} http://products.office.com/en-gb/business/Office
\textsuperscript{18} https://www.google.com/edu/
\textsuperscript{19} http://www.salesforce.com/uk/products/
\textsuperscript{20} https://www.nectar.org.au/virtual-laboratories-1
Platform as a Service (PaaS): The capability provided to the consumer is to deploy onto the cloud infrastructure consumer-created or acquired applications created using programming, languages, libraries, services, and tools supported by the provider. The consumer does not manage or control the underlying cloud infrastructure including network, servers, operating systems, or storage, but has control over the deployed applications and possibly configuration settings for the application-hosting environment.

This is essentially a kit of parts that can be used by developers to simplify the process of building and deploying applications. Examples include facilities for reliably hosting applications at scale that have been written in common programming languages, such as Amazon Elastic Beanstalk\textsuperscript{21}, Microsoft Azure Web Sites\textsuperscript{22} or Google App Engine\textsuperscript{23}. Providers often expose programming interfaces into their own software and services, such as the Google Maps API\textsuperscript{24}, which is widely used to integrate Google Maps into third party sites and services.

For research infrastructures, PaaS is largely manifest in the form of command line or virtual desktop access to a virtual machine which has been specifically tailored with applications and libraries for a given application domain. An example is CloudBioLinux\textsuperscript{25}, a customised version of the popular Ubuntu Linux operating system distribution.

Infrastructure as a Service (IaaS): The capability provided to the consumer is to provision processing, storage, networks, and other fundamental computing resources where the consumer is able to deploy and run arbitrary software, which can include operating systems and applications. The consumer does not manage or control the underlying cloud infrastructure but has control over operating systems, storage, and deployed applications; and possibly limited control of select networking components (e.g., host firewalls).

This is exemplified by the likes of Amazon Web Services\textsuperscript{26} (AWS), Google Compute Engine\textsuperscript{27} and Microsoft Azure Virtual Machines\textsuperscript{28}. Each of these services gives you near instant access to virtual machines hosted in one of the cloud providers’ data centres, pre-loaded with the operating system and often the application software you require.

In the research community there are examples of groups using VMware’s vCloud Director\textsuperscript{29} software to provide IaaS, also OpenNebula\textsuperscript{30}. Groups are increasingly turning to OpenStack\textsuperscript{31}, as described in Annex B.

\textsuperscript{21} http://docs.aws.amazon.com/elasticbeanstalk/latest/dg/Welcome.html
\textsuperscript{22} http://azure.microsoft.com/en-gb/documentation/services/websites/
\textsuperscript{23} https://cloud.google.com/appengine/docs
\textsuperscript{24} https://developers.google.com/maps/
\textsuperscript{25} http://cloudbiolinux.org/
\textsuperscript{26} http://aws.amazon.com/
\textsuperscript{27} https://cloud.google.com/compute/
\textsuperscript{28} http://azure.microsoft.com/en-gb/services/virtual-machines/
\textsuperscript{29} http://www.vmware.com/products/vcloud-suite/
\textsuperscript{30} http://opennebula.org/
\textsuperscript{31} http://www.openstack.org/
A.3 Deployment Models

**Private cloud**: The cloud infrastructure is provisioned for exclusive use by a single organization comprising multiple consumers (e.g., business units). It may be owned, managed, and operated by the organization, a third party, or some combination of them, and it may exist on or off premises.

**Community cloud**: The cloud infrastructure is provisioned for exclusive use by a specific community of consumers from organizations that have shared concerns (e.g., mission, security requirements, policy, and compliance considerations). It may be owned, managed, and operated by one or more of the organizations in the community, a third party, or some combination of them, and it may exist on or off premises.

**Public cloud**: The cloud infrastructure is provisioned for open use by the general public. It may be owned, managed, and operated by a business, academic, or government organization, or some combination of them. It exists on the premises of the cloud provider.

**Hybrid cloud**: The cloud infrastructure is a composition of two or more distinct cloud infrastructures (private, community, or public) that remain unique entities, but are bound together by standardized or proprietary technology that enables data and application portability (e.g., cloud bursting for load balancing between clouds).
Annex B: Cloud Computing for Researchers

Many of the products and services we refer to here are either open source, or available in a free or low cost public cloud platform tier that researchers could readily experiment with.

B.1 Software as a Service

As noted above, software as a service is really another way of saying that the product is delivered as a website or web service. Software as a Service products are typically hosted by the firm which has produced them, and hence not generally available through private or community clouds. We have subdivided this category into two:

Generalised products that have relevance to researchers

Examples here include a wide range of Internet collaboration suites such as Box\(^{32}\), Dropbox\(^{33}\), Google Drive\(^{34}\) and Microsoft OneDrive\(^{35}\). Many of these products have a dual track model that includes a freemium service aimed at consumers, and an enterprise service for paying customers. These tend to have very different terms and conditions, for example consumer terms and conditions often explicitly rule out any warranty whereas enterprise agreements will provide a Service Level Agreement including penalty clauses if SLA commitments are not met. Bespoke enterprise agreements are often created that address customers’ particular concerns, following legal and contractual review.

We will make a special mention here of the Google Apps for Education and Microsoft Office365 collaboration suites, which have been widely taken up by research and education institutions. We can surmise that this is because they offer world-leading facilities at zero cost for the base service. These same services have a significant per-user cost for businesses, but are being widely adopted by firms because of a perception that they significantly undercut the costs of procuring conventional server and storage equipment and then installing and running the closest equivalent software locally - such as Microsoft’s Exchange product. Furthermore, these cloud products are continually being developed and evolved. Google notably take pride in making hundreds of small point releases of Google Apps in a typical twelve month period.

We have also seen a significant amount of interest from researchers in using generalised communication tools such as social media and blogs to disseminate results, to support open practices, and to help stimulate a public debate about their work. Researchers working in scientific computing often exploit products and services aimed

\(^{32}\) https://www.box.com/en_GB/home/
\(^{33}\) https://www.dropbox.com/
\(^{34}\) https://www.google.com/drive/
\(^{35}\) https://onedrive.live.com/about/en-us/
at the developer community such as GitHub\textsuperscript{36}, version control, which offers both a freemium consumer service and an enterprise service.

**Products that are specifically targeted at researchers**

In parallel to research adoption of the sort of generalised services listed above, a number of products have emerged that specifically target researchers as a community. These include:

- Professional networking sites such as Mendeley\textsuperscript{37}, Academia.edu\textsuperscript{38} and ResearchGate\textsuperscript{39}.
- Specialised facilities for sharing code and data such as figshare.com\textsuperscript{40}.
- Open access journals, repositories and pre-print archives such as arXiv.org\textsuperscript{41}, PLOS\textsuperscript{42}, institutional repositories and sites from traditional publishers such as Elsevier's ScienceDirect\textsuperscript{43}.
- Equipment catalogues such as equipment.data.ac.uk\textsuperscript{44} and Kit-Catalogue\textsuperscript{45} from Jisc and EPSRC.
- Lab management software such as Quartzy\textsuperscript{46} – a free service which coincidentally links users of lab equipment with vendors to reduce the friction of replacing consumables and equipment more generally.
- Lab tests “as a service” – send off your sample to be processed and download the results.
- Independent Software Vendors (ISVs) of scientific computing software and major hardware providers are in many cases starting to offer their own packages through their own cloud service – e.g. the Atos Extreme Factory\textsuperscript{47}.

**B.2 Platform as a Service**

As noted above, it has become increasingly common for cloud providers to expose the underpinnings of their internal services for other developers to build upon. These tend to be promoted first and foremost as facilities for helping developers to build high volume websites - such as load balancing and highly performance/available databases. However, cloud providers also often expose "microservices" that are of particular relevance to

\begin{footnotesize}
36 http://github.com
37 https://www.mendeley.com/
38 https://www.academia.edu/
39 http://www.researchgate.net/
40 http://figshare.com/
41 http://arxiv.org/
42 https://www.plos.org/
43 http://www.sciencedirect.com/
44 http://equipment.data.ac.uk/
45 http://www.kit-catalogue.com/
46 https://www.quartzy.com/
47 http://www.bull.com/extreme-factory
\end{footnotesize}
researchers – e.g. machine learning and data processing facilities. We have picked out some examples below. It is important to note here that many of the key platform tier products are either open source or available in an open source equivalent, facilitating their use in private and community clouds.

Relational databases

Most of the major public cloud providers have some form of relational database platform service. These are usually compatible with existing widely deployed products – e.g. Amazon RDS\(^{48}\) can be used in place of MySQL simply by changing the Database Source Name (DSN) your code connects to.

NoSQL

Cloud providers often offer so called NoSQL services, which are typically non-relational (schemaless) approaches to storing and manipulating large volumes of information. Examples here include Google Cloud Datastore\(^{49}\), Amazon DynamoDB\(^{50}\), several Azure NoSQL services and the open source Cassandra\(^{51}\), CouchDB\(^{52}\) and MongoDB\(^{53}\) products.

Object stores

Software built to use a cloud model has tended to avoid having permanently mounted shared filesystems such as we might see in an HPC cluster. Instead, it is common practice to use object stores such as Amazon S3\(^{54}\), Google Cloud Storage\(^{55}\), Azure Blob Storage\(^{56}\) or open source packages like Redis\(^{57}\). Researchers might be more familiar with using more niche (yet very well established in the research community) object store software like the open source iRods\(^{58}\), which is not yet offered directly as a service by the major cloud providers.

“Big data”

The gold standard software for handling big data is the Apache Hadoop\(^{59}\) project, an open source suite which gives roughly equivalent capability to Google’s proprietary BigTable algorithm via its HBase subsystem. Hadoop is widely available from cloud providers as a hosted service, and also often sold to enterprise and institutions as a big data handling solution – e.g. via a packaged appliance/cluster. Subsequent to the success of Hadoop, Google

\(^{48}\) http://aws.amazon.com/rds/
\(^{49}\) https://cloud.google.com/datastore/docs/concepts/overview
\(^{50}\) http://aws.amazon.com/documentation/dynamodb/
\(^{51}\) http://cassandra.apache.org/
\(^{52}\) http://couchdb.apache.org/
\(^{53}\) https://www.mongodb.org/
\(^{54}\) http://aws.amazon.com/s3/
\(^{55}\) https://cloud.google.com/storage/
\(^{56}\) http://azure.microsoft.com/storage/
\(^{57}\) http://redis.io/
\(^{58}\) http://irods.org/
\(^{59}\) https://hadoop.apache.org/
opened up its own internal service as the BigQuery\textsuperscript{60} product. It is becoming common for cloud providers to provide “big data as a service” using Hadoop or their own proprietary technology, e.g. Amazon offer their Elastic MapReduce\textsuperscript{61} product.

**Machine learning**

Machine learning is generally defined in terms of “teaching” the algorithm to recognise a given target using training datasets. For example, Google recently used a cache of cat videos on YouTube to create a machine learning model that was reliably able to identify cats in videos – and work has now moved on more complex concepts such as describing the contents of a photograph. Machine learning tools such as Amazon Machine Learning\textsuperscript{62}, Google Prediction API\textsuperscript{63}, Azure Machine Learning Studio\textsuperscript{64} and Apache Mahout\textsuperscript{65} are generally available. Mahout is open source, but not yet offered as a platform service by the major cloud providers, who have their own alternatives.

It is important to note that many platform services expose Application Programming Interfaces (APIs) that are unique to that service. Whilst this is not universally true (e.g. Amazon RDS “looks like” MySQL), there is a potentially significant effort required to migrate from one provider’s platform service to another. Furthermore, some of the associated data, such as a machine learning model, may not be readily exported to another provider. Therefore switching costs will include the regeneration of that dataset.

**B.3 Infrastructure as a Service**

This is the cloud tier which researchers in scientific computing may be best equipped to engage with directly, and indeed many researchers are using private, community or hybrid clouds today without perhaps realising it – for example CERN has moved its entire compute infrastructure over to the OpenStack\textsuperscript{66} open source cloud software.

Infrastructure as a Service typically equates to running up virtual machines (VMs) on the cloud provider’s shared (multi tenant) infrastructure but it equally applies to storage and networking configuration. Cloud providers have various strategies for preventing these virtual machines from interfering with one another, such as security groups to provide IP address level access controls. Whilst there are a number of virtual machine image formats used, tools are also available to convert from one to another, and in most cases it is possible to take a virtual machine image from one provider and convert it to the format used by another – e.g. from a VMware VMDK disk image or an OpenStack QCOW2 image to an Amazon Machine Image (AMI).

\textsuperscript{60} https://cloud.google.com/bigquery/
\textsuperscript{61} http://aws.amazon.com/elasticmapreduce/
\textsuperscript{62} http://aws.amazon.com/machine-learning/
\textsuperscript{63} https://cloud.google.com/prediction/
\textsuperscript{64} https://studio.azureml.net/
\textsuperscript{65} http://mahout.apache.org/
\textsuperscript{66} http://www.openstack.org/
Orchestration

Infrastructure as a Service is very well established, with a wide range of orchestration tools from cloud providers and the open source and commercial software communities. At one extreme these permit the user to bring up a complete suite of VMs featuring different applications and configurations, to create virtual data centres. At another extreme we might simply be interested in ensuring that certain software dependencies are installed on a VM. Orchestration examples from cloud providers include Amazon’s CloudFormation\(^{67}\), Google Cloud Deployment Manager\(^{68}\) and Azure Automation\(^{69}\). There are also provider-agnostic orchestration tools such as OpenNebula\(^{70}\) (AWS, Azure and private/hybrid clouds).

Canned VMs

One of the more daunting aspects of supporting scientific computing can be the sheer range of software packages in use by an institution’s researchers. Most major cloud providers offer a marketplace of ready-made VMs with commonly used software pre-installed on them, such as Amazon’s AWS Marketplace\(^{71}\) or Microsoft’s VM Depot\(^{72}\). This has the potential to be a huge time saver for both the researcher and the scientific computing team, e.g. for an application with complex dependencies such as the Galaxy\(^{73}\) bioinformatics suite.

However, the researcher may need to be extremely careful about tracking software versions when using readymade VMs, to be sure of being able to reproduce results. For example, it may be advisable to archive the VM used for a particular piece of work, just in case this turns out to no longer be available through the cloud provider at a later date when it is necessary to repeat the work.

Bring Your Own License

It is becoming increasingly common to find both open source and commercial off the shelf scientific computing software available packaged as VMs, e.g. ANSYS provide their FLUENT CFD software in this way through the AWS Marketplace. Commercial software is typically provided under a Bring Your Own License model, where the end user is expected to already be a licensed customer - perhaps with a negotiated license extension to use cloud facilities. Without the license key or a connection to a license server, the software is useless. In some cases the Independent Software Vendor (ISV) operates their own cloud based license server for real time checking of license entitlement against concurrent usage.

Quasi-HPC facilities

---


\(^{68}\) [https://cloud.google.com/deployment-manager/](https://cloud.google.com/deployment-manager/)


\(^{70}\) [http://opennebula.org/](http://opennebula.org/)

\(^{71}\) [https://vmdepot.msopentech.com/List/index](https://vmdepot.msopentech.com/List/index)

\(^{72}\) [https://vmdepot.msopentech.com](https://vmdepot.msopentech.com)

\(^{73}\) [https://galaxyproject.org/](https://galaxyproject.org/)
Of course many researchers are accustomed to running their codes on traditional HPC clusters, and cloud providers have recognised this by providing tools that let the researcher conveniently bring up a large number of VMs with dedicated login nodes and shared storage in the typical HPC cluster model. Examples of these include the Azure HPC Cluster Service\(^74\) and Amazon’s cfncluster tool\(^75\). These “quasi clusters” often do not have the sort of performant interconnect and high performance parallel filesystem that we might see on a true HPC cluster – although Microsoft now offer compute nodes with Infiniband interconnect as part of Azure.

For many classes of compute workload this may not be a blocker (e.g. embarrassingly parallel bioinformatics jobs), but further work is still required with commonly used codes to explore performance aspects. Some cloud providers have recognised that there is sufficient demand for high performance interconnects and are making specialist hardware available such as extra large nodes, low latency interconnects, GPUs and even FPGAs.

**Virtualization, containers and Docker**

In an ideal world, it would be possible to simply pick up an application and move it and its dependencies to the compute platform of choice. This is something that the scientific computing community has often addressed by statically linking executables to the libraries that they depend on, but in today’s increasingly complex data processing environment this approach has its limitations. Virtualization at the machine level has made it possible to build the desired environment, e.g. on the researcher’s laptop, and then take a snapshot of it to run on the compute platform of choice. However, the wide range of machine images and hypervisors makes it difficult to generalise this approach – what is needed is some kind of packaging standard. A further wrinkle is that hypervisors typically add an unwelcome overhead to compute and data intensive workloads.

The Docker\(^76\) project attempts to solve these problems using via operating system level virtualization approach, building on Linux kernel features to provide portable provider agnostic “containers” that encapsulate applications and their dependencies whilst isolating them from each other and improving on the performance of a hypervisor approach to virtualization. This is actually not a new approach, and Docker builds on ideas that may be familiar from Solaris Zones and mainframe operating systems. Docker containers are inherently portable. Support for Docker has been forthcoming from Amazon\(^77\), Google\(^78\) and Microsoft\(^79\) amongst others. Whilst there are other competing technologies, Docker has reached critical mass in terms of mind share and is widely used within the Internet industry e.g. eBay, Yelp, Spotify, Yandex and Baidu. The formation of the Open Container Project\(^80\) in summer 2015, with broad industry support, suggests that application virtualisation is set to rapidly become the norm.

---

75 http://aws.amazon.com/hpc/cfncluster/
76 http://docker.io
77 http://docs.aws.amazon.com/AmazonECS/latest/developerguide/docker-basics.html
78 https://cloud.google.com/compute/docs/containers
80 https://www.opencontainers.org/
Annex C: Trust & public cloud

Post-Snowden, we have increasingly started to cast a critical eye on services provided by US owned tech firms - have they colluded with the FBI, CIA or NSA to build in back doors to their products? The UK Government has provided\(^85\) a convenient set of Cloud Security Principles that researchers and institutions can use to establish whether a provider has taken adequate care to protect their data. Leading public cloud providers like Amazon\(^82\) and Microsoft\(^83\) have provided their own compliance statements regarding the Cloud Security Principles.

Public cloud providers would also note that it is perfectly possible to create an insecure or unreliable cloud service simply by not following best practice, just as it was always possible to create an insecure or unreliable in-house service. Amazon codify this through their statement about shared responsibilities, shown in the figure below.

![Amazon Shared Responsibility Model](image)

**Figure 2. Amazon shared responsibility model**

For researchers and institutions operating in the UK and the greater European Economic Area, there are particular issues around data sets that it has been stipulated may not leave the country, or leave Europe as a whole. An example of this would be genome data gathered as part of Genomics England’s 100,000 Genomes project - which must be kept in England\(^84\).

Conversely, the EU-SafeHarbor Agreement\(^85\) which creates a managed process governing the controlled release of European data to the United States, is currently being challenged in the European Court of Justice by Austrian law student Maximillian Schrems, who alleges that it contravenes EU Data Protection legislation\(^86\). If the Schrems case is successful, EU-SafeHarbor will be struck from the books and US service providers may be forced to open European facilities or prevented from operating in Europe.

---

81 https://www.gov.uk/government/publications/cloud-service-security-principles
84 http://www.genomicsengland.co.uk/the-100000-genomes-project/faqs/data-faqs/
We might draw the conclusion from Snowden that all forms of digital technology have been or will be infiltrated by national actors, but it is also the case that the major cloud providers have been working tirelessly to reduce the attack surface - e.g. by encrypting links between data centres, using two factor authentication, and encrypting data in transit and at rest. For sound business reasons, cloud providers are trying to ensure that an auditable and judicial process is followed when providing information about users or their data to authorities, and to frustrate “trawling” efforts. For example, Microsoft mounted a legal challenge around the release of data from their Dublin Azure data centre to the US government\(^{87}\), and Google publish a regular Transparency Report\(^{88}\) quantifying government requests for data.

To help establish some norms around use of public cloud globally we have provided several examples below of case studies where cloud technologies being used for sensitive applications including personal data, government data and in cases where intellectual property is a key consideration:

- “The Financial Industry Regulatory Authority (FINRA) in the US is using AWS to analyze and store approximately 30 billion market events every day, saving some $10m-$20m through the move to the cloud” - FINRA\(^{89}\)

- “In the past, a simple question about genetics linked to a medical condition might take hours, or even days, to execute. By leveraging Google Cloud Platform, the analysis of 1,000 patients’ genomic data, across 218 diseases, generates near real-time results” – Northrup Grumman\(^{90}\)

- “The Philips HealthSuite digital platform analyzes and stores 15 PB of patient data gathered from 390 million imaging studies, medical records, and patient inputs to provide healthcare providers with actionable data, which they can use to directly impact patient care” – Philips\(^{91}\)

- “Mount Sinai and their collaborators at Station X are mining the more than 2,000 breast and ovarian tumor and germline DNA sequences (100TB data) generated by The Cancer Genome Atlas Consortium” – Mount Sinai Medical Centre\(^{92}\)

- “Instead of having to spend in the order of £50,000 per year on storage, we can expand our cloud storage or buy some tier-three storage instead. That’s an order of magnitude cheaper - we can literally knock a zero off that sum when we need to expand” – Homerton Hospital\(^{93}\)

From a UK perspective we would also note that Google Apps is currently being rolled out across HM Revenue & Customs\(^{94}\), and the UK Supreme Court has moved to Office365\(^{95}\).

---


\(^{88}\) [http://www.google.com/transparencyreport/](http://www.google.com/transparencyreport/)


\(^{90}\) [http://googlecloudplatform.blogspot.co.uk/2015/03/personalized-medicine-with-northrop-grumman-and-google-cloud-platform.html](http://googlecloudplatform.blogspot.co.uk/2015/03/personalized-medicine-with-northrop-grumman-and-google-cloud-platform.html)


Annex D: NeI and cloud

D.1 Public investment in e-Infrastructure

Investments by BIS, the Research Councils and HEIs in 2011-12 (£160M), 2012-2013 (£189M) and 2014-15 (£257M) have resulted in core elements of the national e-Infrastructure being put in place.

2011-2012

Investments were made in core HPC and Networking infrastructure. In addition investments were made in the Authentication Infrastructure Moonshot (now known as Jisc Assent). These investments are listed in Table 1 below.

Table 4: 2011-2012 HPC Investments

<table>
<thead>
<tr>
<th>HPC Project</th>
<th>RC</th>
<th>Amount/£M</th>
</tr>
</thead>
<tbody>
<tr>
<td>National Service</td>
<td>EPSRC, NERC</td>
<td>43</td>
</tr>
<tr>
<td>Hartree Centre</td>
<td>STFC</td>
<td>30</td>
</tr>
<tr>
<td>DIRAC</td>
<td>STFC</td>
<td>15</td>
</tr>
<tr>
<td>GRIDPP</td>
<td>STFC</td>
<td>3</td>
</tr>
<tr>
<td>The Genome Analysis Centre (TGAC)</td>
<td>BBSRC</td>
<td>8</td>
</tr>
<tr>
<td>Monsoon</td>
<td>NERC/Met Office</td>
<td>1</td>
</tr>
<tr>
<td>JASMIN2 &amp; CEMS</td>
<td>NERC, &amp; UKSA</td>
<td>7.75</td>
</tr>
<tr>
<td>Regional Centres: N8, SES5, MID+,</td>
<td>EPSRC</td>
<td>10</td>
</tr>
</tbody>
</table>

### HPC Project

<table>
<thead>
<tr>
<th>HPC Project</th>
<th>RC</th>
<th>Amount/£M</th>
</tr>
</thead>
<tbody>
<tr>
<td>WeSt</td>
<td></td>
<td></td>
</tr>
<tr>
<td>JANET Network and Authentication (Moonshot)</td>
<td>Jisc</td>
<td>31</td>
</tr>
<tr>
<td>HPC Data Storage</td>
<td>EPSRC, STFC</td>
<td>15</td>
</tr>
</tbody>
</table>

### Table 5: Big Data Investments 2012-2013

<table>
<thead>
<tr>
<th>Big Data Project</th>
<th>RC</th>
<th>Amount/£M</th>
</tr>
</thead>
<tbody>
<tr>
<td>Digital transformations in arts and humanities</td>
<td>AHRC</td>
<td>8</td>
</tr>
<tr>
<td>E-infrastructure for biosciences</td>
<td>BBSRC</td>
<td>13</td>
</tr>
<tr>
<td>Research data facility and software Development</td>
<td>EPSRC</td>
<td>8</td>
</tr>
<tr>
<td>Administrative data centres</td>
<td>ESRC</td>
<td>36</td>
</tr>
<tr>
<td>Understanding populations</td>
<td>ESRC</td>
<td>12</td>
</tr>
<tr>
<td>Business datasafe</td>
<td>ESRC</td>
<td>14</td>
</tr>
<tr>
<td>Biomedical informatics</td>
<td>MRC</td>
<td>55</td>
</tr>
</tbody>
</table>

2012-2013

Big Data projects using funds announced by the Government in December 2012 were funded at this time. Major Awards have been made to 18 centres in the UK, 16 of whom are HEIs. These awards are listed in Table 2. The pre-eminent role of HEIs in managing and providing national and Large Specialist data and compute services to UK academia is emphasised by these awards.
Further investments were made as follows:

- The Medical Research Council (MRC) will invest £50 million in bioinformatics, which uses many areas of computer science, statistics, mathematics and engineering to process biological data. These include £19M to the Farr Institute of Health Informatics Research at nodes, London, Manchester, Wales, Scotland; and the £32M Medical Bioinformatics initiative to fund 5 projects; eMedLab (UCL Partners-Crick-Sanger-EBI), The MRC Consortium for Medical Microbial Bioinformatics, Leeds MRC Medical Bioinformatics Centre and the MRC/UVRI Medical Informatics Centre.
- The Arts and Humanities Research Council (AHRC) invested £4 million in 21 new open data projects. They will make large data sets that ordinarily only academics would have access to accessible to the general public.
- The Economic and Social Research Council (ESRC) has invested £14 million in 4 new research centres at Essex, Glasgow, UCL and Leeds Universities, as well as a further £5M investment in the Administrative Data Service at Essex University. The centres will make data from private sector organisations and local government accessible to researchers investigating anything from transport to obesity. At present the data is being collected by these organisations, but is not being used for research purposes. The programme is divided up into 3 phases: Phase 1 was set up to get information from government departments and set up Administrative Data Research Network; Phase 2: set up Business and Local Government Data Research Centres; collect and set up Phase 3 Social Media and Third Sector data.
- The Natural Environment Research Council (NERC) has invested £4.6 million of funding for 24 projects to help the UK research community take advantage of existing environmental data.
- The Engineering and Physical Sciences Research Council (EPSRC) have invested £8M in the Research Data Facility (RDF) which is operated by EPCC.

The RDF is designed to provide research data management and data analysis services for ALL RCUK researchers. Access will be governed by a peer review mechanism. Early projects include the hosting of the DiRAC Code Benchmarking Project.

<table>
<thead>
<tr>
<th>Big Data Project</th>
<th>RC</th>
<th>Amount/£M</th>
</tr>
</thead>
<tbody>
<tr>
<td>NERC Environmental Big Data Initiative</td>
<td>NERC</td>
<td>13</td>
</tr>
<tr>
<td>Square Kilometre Array</td>
<td>STFC</td>
<td>11</td>
</tr>
<tr>
<td>Energy Efficiency Computing Hartree Centre</td>
<td>STFC</td>
<td>19</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>189</td>
</tr>
</tbody>
</table>
Other data infrastructure investments include the

- European Bioinformatics Institute at Hinxton (Cambridge) which received £75M from BBSRC,
- the Open Data Institute, a not for-profit funded by Innovation UK (£10M over 5 years, subject to industry investment) and by industry is a big data undertaking dedicated to providing open access to data from across the public sector in order to enable industrial and academic exploitation.
- In 2012, the Clinical Practice Research Datalink, a £60 million service funded by the MHRA and the National Institute for Health Research, was established to provide patient data for medical research.
- The Government earmarked £100 million for the NHS to sequence the DNA of up to 100,000 patients with cancer and rare diseases, which will include the development of appropriate data infrastructure (NHS).

2014-2015

Three major investments dominated this period

1. Centre for Cognitive Computing at the Hartree Centre. This was funded at the £115M level with a further £230M from IBM
2. A 10 Pflop Supercomputer for the Met Office (£100M)
3. Alan Turing Centre for Data Science (£42M)

In addition it was announced that a further £100M would be made available to the SKA Project as part of Big Data Investments.

D.2 NeI, cloud technology & the access agenda

At present the NeI projects are largely hidden from the general research community and some would say are largely hidden from even their intended user base.

HPC and HTC are still regarded as difficult tools to use by researchers, even though they are probably no more difficult to use than a piece of lab equipment.

How do we then make HPC, Data Intensive Computing and HTC be regarded as part of their basic research laboratory?

This requires researchers to view these resources as
• Viewable. They are manifested tangibly, whether on a desktop or a mobile device
  • Easy to interface with so that users can
    o Submit workflows
    o Construct workflows
    o See what resources are available
    o Check what resources they have used
  • Needing only ONE user identity

The diagrams above and below gives an example of how this works. As far as the user is concerned directly interfacing is now not needed. What becomes important now are the Cloud Services described earlier in this report.

The need to interact directly with the compute and data resources is removed and is replaced by a set of generic functions in a workflow that allow the user in effect to use the same workflow on a multiplicity of systems.

It is this that removes a barrier to usage by given user the on-demand self-service they would like to use.
An example in the UK of a functioning Cloud is JASMIN2\textsuperscript{96}. JASMIN2 supports the data analysis requirements of the UK and European climate and earth system modelling community. It consists of multi-Petabyte fast storage co-located with data analysis computing facilities, with satellite installations at Bristol, Leeds and Reading Universities.

JASMIN2 is a successful Cloud deployment that services a key UK research community.

\textsuperscript{96} http://jasmin.ac.uk/