Report of 2014 UK National e-Infrastructure Survey (HEIs and Research Institutes)

Contents

Executive Summary 2

1. Progress Report 6

1.1 Common Authentication, Authorisation and Access Methods 6

1.2 Connectivity and Data Sharing 6

1.3 e-Infrastructure Resource Inventory 7

1.4 Scientific Computing Software Expertise within Institutions 7

1.5 A Shared Approach to Training and Support 7

1.6 Funding for Code Development and Maintenance 8

2. Capital Investments in Big Data Facilities 10

2.1 Particular Issues around Access 12

3. Who are the Ne-I's Service Providers 14

3.1 HEIs 14

3.2 National and International "Centres" 15

4. The National E-Infrastructure and Industry 17

5. Conclusions and Recommendations 19

Appendix A: List of Surveyors and Acknowledgements 21

Appendix B: Who Will Receive this Survey 21

Appendix C: List of Respondents 21

Large and Specialist Services: 21

Service Management Responses Only: 21

HEIs: 22

Appendix D – Service Management: The Survey Questions 23

Appendix E – Service Management: Summary of the Survey Data 25

Appendix F – Service Management: Full Break Down of Survey Data 27

Appendix G – Hardware: The Survey Questions 42

Appendix H – Hardware: Summary of the Survey Data 44

Written Responses 52

The EMBL European Bioinformatics Institute 52

The Farr Institutes 52

Admin Data Service and Admin Data Research Centres 53
Appendix I – Use of software for scientific computing in the UK 55

National e-Infrastructure Survey 2014

Jeremy Yates (Chair, PDG) and Martin Hailton (JISC)
Harpreet Dhanoa (DIRAC) and Clare Jenner (DiRAC)
Neil Chue Hong and Simon Hettrick (Software Sustainability Institute)
Oz Parchment (HPC-SIG), Alan Real (HPC-SIG), Andrew Richards (Regional HPC), Simon Burbidge (HPCSIG)

Executive Summary

This is the second Survey of the National e-Infrastructure (Ne-I) that is located in the UK's Higher Education Institutes (HEIs) and Research Institutes. We have been able to perform a comprehensive survey of the state of the new National E-Infrastructure. An in depth Survey of Service Management was conducted as well as Hardware and Software Surveys.

We have seen significant investment in the core of the JANET network and in e-Infrastructure facilities connected to it. Campus network infrastructure is a critical part of an institution's capacity to exploit e-Infrastructure investments. There are bottlenecks appearing in local campus networking that do now need to be addressed.

Recommendation: The connections of Ne-I Service providers to the SJ6 backbone be evaluated and if necessary upgraded or separate links be provided.

Recommendation: Internal investment by institutions is required to ensure that internal campus networks remain fit for purpose. This needs to be communicated to HEIs via their Pro-VCs for Research.

Significant new capability has come online this year, with the deployment of ARCHER and the new Big Data Services greatly expanding the power of the Ne-I. We note that Services such as DiRAC are now more than half way through their operating lifetime and will require renewal in 2015-2016.

Recommendation: A long term capital plan is required to ensure the future productivity of the Ne-I. This should be co-ordinated and carry on momentum established to create a holistic e-Infrastructure eco-system for the UK. Being able to plan will greatly increase the efficacy and efficiency of our systems and release resources for added value activities such as software engineering. Greater co-ordination has the potential to establish deeper and more valuable partnerships with the vendor community.

Progress has been made towards producing a common single-sign-on and data security infrastructure. This is often known as Authentication, Authorisation and Accounting Infrastructure (AAAI). RCUK has set up an Access and Security Working Group to co-

ordinate efforts in these areas. There has been widespread interest in the Moonshot technology trial, but this in itself only addresses part of the problem - the Authentication step of reducing the friction of access to e-Infrastructure services. There are a number of further issues around authorisation and accounting, information assurance and the legal and contractual framework within which e-Infrastructure services are shared between organisations – potentially leading to a single e-Infrastructure service portal.

Recommendation: JANET (JISC) is well placed to continue coordinating efforts in this area, building upon existing work wherever possible.

Recommendation: DiRAC, GRIDPP and JANET work together to produce a prototype that allows these Ne-I Projects to share resources. This tests the single-sign-on capability of the proposed common AAAI infrastructure

Recommendation: EMEDLAB, FARR Institutes, ARDC, EBI/ELIXIR and JANET work together to produce a secure AAAI infrastructure that protects the security of the sensitive *people* data. This tests the data security aspect of the proposed common AAAI infrastructure.

Historically HPC focussed e-Infrastructure services have tended to work on the basis that their users are predominantly writing or working with parallelised codes. However it is increasingly the case that the majority of users simply wish to run much simpler (though sometimes embarrassingly parallel) codes on a "bigger" machine – with more RAM per core, storage, and so on. Cloud computing technologies such as OpenStack and Linux Containers / Docker can potentially play a major part here, with a standard library of virtual machine appliances preconfigured for common scientific computing workflows.

Recommendation: JASMIN, Sanger Institute, GRIDPP, SKA and DiRAC should plan to work together to explore the practicalities of this approach and show that resources within Research Domains can be configured into effective and efficient private clouds that allow researchers to run their workflows easily on a domain private cloud or on resources in another part of the Ne-I.

There is a common perception that commodity public cloud computing services are not cost effective for compute and data intensive applications. Through the recent JANET cloud computing service agreements good relations have been built up with the major service providers and it is now a good time to expand the range of service providers to the National e-Infrastructure.

Recommendation: JANET to take this work forward, in collaboration with major e-Infrastructure service users and public cloud providers. However the costs and benefits of such access will have to be carefully measured and it is unlikely to be a solution for many of our problem sets in the next 2-3 years. Public Cloud Providers and Ne-I providers should exchange information concerning the problem types and sizes in the Ne-I, so that expectations are not unduly raised and that methodologies are built up that allow effective and economic use of Public Clouds.

Whilst concerted programmes of work exist around large and specialist facilities, e-Infrastructure provision in HEIs has not had the same level of central support and coordination.

It is clear from the survey results that HEIs provide most of the Ne-I services to our communities and that many HEIs are not reaping the full benefits from their e-Infrastructure investments due to the overhead involved in "keeping the show on the road". Survey respondents tell us that this often comes close to (or exceeds) the available FTE count. We recommend that the funders consider potential inducements to improve this situation.

Recommendation: The e-Infrastructure Leadership Group is encouraged to consider potential approaches such as greater regional collaboration supported by an element of matched funding for the HEIs' e-Infrastructure investments. This should be highlighted to Pro-VCs for Research.

It is clear we have several National Centres that underpin our Ne-I. We also have significant new hardware R&D capability provided by the Square Kilometre Array (SKA) IT Open Architecture Laboratory and the Hartree Centre Energy Efficient Computing Unit. This provides an exciting new possibility for the UK as it allows the UK to help develop HPC and Big Data Technologies and systems, as well as develop software products to utilise these technologies.

Recommendation: There is now the opportunity to formally recognise and define the contribution of the National "Centres" to the Ne-I and to make sure they are adequately funded to carry out their individual missions.

Scientific software remains a critical part of the UK's e--Infrastructure. Compared to the results of the 2013 survey, the responses from the 2014 survey show that the main issues and concerns remain. As recognition and provision become more widespread, there is convergence on the importance of *basic software engineering training for researchers*, *research software engineer career paths*, and *sustaining a critical mass of software expertise*.

There is a clear increase in the awareness of the need to provide support for Software, Software Developers and Training. The obvious standouts were *career paths for developers of scientific software* and *basic software engineering training for our user-researchers*. For software issues, scientific consortia highlighted optimisation methods, recognition of software as a research object / credit for software and recognition of software planning / sustainability in grants by investigators and reviewers, whereas institutions highlighted, reproducibility / correctness of results, sustainability of software, more robust Linux installation processes, general training for staff and postgraduates, licensing, software as a research output and new technology expertise.

For training, scientific consortia felt that the following topics were underprovided: Basic software engineering skills; using clusters / parallel and distributed programming; Data analysis techniques; choosing and using software appropriately and Bid writing, which includes software. Institutions highlighted: Basic software engineering skill; how computational techniques can be applied as research tools; data analysis; data curation and management; courses tailored for X-informatics subjects and parallel programming.

It was felt that training is still undervalued as an activity and is expected to be provided for free. There needs to be a clear Branscomb pyramid in terms of Training and Support.

Recommendation: Funders should make applicants aware that it is permissible to apply for research software engineer time on grants, and that it is appropriate to class these as research staff on grants where the work to be carried out involves a significant research and development aspect.

Recommendation: The work to raise the profile of the research software engineer should continue - in particular the role, value and potential career paths should be highlighted into submissions to the ELC and HEIs. The value of dedicated developer support at HEIs should also be highlighted to Pro-VCs for research and Directors of Research, as well as to successful funding models.

Recommendation: Training materials and courses should be easily available. A service to host materials and advertise courses and to assign courses to particular levels of skill and knowledge would be very helpful in making sure users get appropriate training and are able to progress easily from level to level. HEIs and Research Domains should be encouraged to share what materials and resources they can with each other. Training has to be provided for trainers.

Recommendation: A Training Framework, possibly on an on-line Marketplace, needs to be created to allow our user base to find out and access the right training needed for their project.

The creation of on-ramp centres under the auspices of InnovateUK (TSB) should allow SMEs to access Ne-I resources. Much of the work outlined in this report is designed to remove barriers to SMEs using Ne-I resources. The creation by InnovateUK (Technology Strategy Board (TSB)) of the e-I Special Interest Group has brought Ne-I providers and SMEs and Primes together so that needs and requirements can be discussed and enumerated.

Recommendation: A process is put in place to make sure the on-ramp centres and the Ne-I work together so that SMEs can actually access Ne-I resources. This should include schemes to induce SMEs to use Ne-I infrastructure.

1. Progress Report

The 2013 e-Infrastructure Survey[1] made a number of recommendations that have subsequently been acted on by funders, policymakers and the e-Infrastructure community. Progress against these recommendations is set out below.

1.1 Common Authentication, Authorisation and Access Methods

The community has made significant progress towards a common authentication substrate, with 35 universities, colleges and research institutes piloting the IETF standards track Moonshot technology. Moonshot is a JANET initiative in partnership with GÉANT and others, to develop a single unifying technology for extending the benefits of federated identity to services beyond the web, including cloud infrastructures, high performance computing, grid infrastructures and other services such as email. Moonshot builds upon the underlying mechanisms used in eduroam and the UK Access Management Federation (Shibboleth, SAML, RADIUS etc).

The JASMIN service, run for the climate and earth sciences community by STFC, has developed a model for resource allocation akin to a commodity public cloud provider, using virtualisation to offer researchers a choice of machine images pre-loaded with relevant scientific computing software. JASMIN, GRIDPP, DiRAC and the SKA project are also exploring a "Bring Your Own Workflow" approach, whereby the researcher can develop a virtual machine image which they then upload to a compute cluster for execution.

Significant developments are also under way in access management and virtualisation through the SKA's Open Architecture Laboratory, the Hartree Centre's Energy Efficient Computing project, and the work under way to create the Farr Institute for Health Informatics Research. The e-Infrastructure Project Directors Group have an aspiration to (i) harmonise this "AAAI" work across the GRIDPP and DiRAC services, and (ii) to develop a standardised infrastructure for AAAI and a framework for assured data exchange for Administrative Data Centres, Medical Bioinformatics and Farr Institute partners.

1.2 Connectivity and Data Sharing

The launch of JANET6 in autumn 2013 saw significant capacity and connectivity improvements. JANET6 provides a long term stable platform upon which to build, with strategic investment in the underlying fibre network for a ten year period and planned equipment refreshes built into the service provision. Aurora2, the National Dark Fibre Infrastructure Service, has just been launched. Recognising the increasing requirement for industrial access to e-Infrastructure facilities and services, the JANET Reach scheme was launched in spring 2014 to pilot industrial connectivity to JANET for R&D collaboration. The first round of JANET Reach applications is currently being evaluated.

[1] http://www.clms.ucl.ac.uk/sites/default/files/NationalEinfrastructureSurvey%20Project%20Directors%20Group%20March%202013.pdf

Funding from BIS also made it possible to improve JANET connectivity for a number of strategic research facilities, including connections to Norwich Research Park and the Hinxton Genome Campus (BBSRC) and the Met Office at Exeter (NERC). Connectivity to the Francis Crick Institute has been planned and is awaiting completion of the building project. JANET are also facilitating the Shared Higher Education Data Centre procurement on behalf of a wide consortium of Higher Education Institutions and Research Institutes including UCL, LSE, Kings' College, QMUL, the Francis Crick Institute, the Wellcome Trust Sanger Institute and others.

1.3 e-Infrastructure Resource Inventory

The National e-Infrastructure Survey continues to be an effective means of gathering information from the community on both large and specialist facilities and also HEIs. Responses were received from 35 of the 38 HEIs participating in HPC-SIG, and 17 large and specialist facilities.

It was necessary to accelerate the survey in order to meet tight deadlines and several respondents felt that more time should have been allowed. This has been noted and will be reflected in the timing of the 2015 survey, which will be formally conducted by JISC on behalf of RCUK and the e-Infrastructure community.

The RCUK National e-Infrastructure Group is working with the InnovateUK (TSB) to take forward and make accessible the inventory of facilities produced as part of the e-Infrastructure survey. This will ensure that the public investment in e-Infrastructure is exploited to its fullest by both the education sector and industry.

1.4 Scientific Computing Software Expertise within Institutions

The survey results indicate that institutions typically have very small teams working in support of research / scientific computing – often as little as one or two FTE even for a major HPC facility with a large number of users. Respondents indicated that a large proportion of staff time in HEIs in particular is necessarily devoted to system management issues. Given the small team sizes that are typical at present, this reduces the staffing resource available for training and outreach.

Where institutions do not have well established and potentially self-supporting research groups, significant gaps have been identified in terms of training and support for off-the-shelf software, assistance for researchers in terms of bootstrapping, software development best practice (such as version control and continuous integration testing) and re-use of existing code libraries. Linux skills in particular were identified as a critical development requirement for researchers working in scientific computing.

1.5 A Shared Approach to Training and Support

A new e-Infrastructure Trainers Special Interest Group had been formed under the auspices of the Software Sustainability Institute. This held a landmark workshop for e-Infrastructure trainers in August 2013, which brought together 50 trainers from across

the UK to work on plans for improving e-Infrastructure training and to determine the benefits of creating a formal training community. Key recommendations from the workshop[2] follow:

- The SIG will provide a single point of contact for the trainers in the community, which is vital for dissemination and lobbying;
- Collection of success metrics from e-Infrastructure training activities, to demonstrate the value to institutions and funders of investing in them;
- Creation of a single access point to training would be a significant improvement over the current spread of resources across many different organisations;
- By working together, the training community will increase the quality and visibility of resources, making it easier for researchers to acquire the skills needed to harness e-infrastructure for their research.

A large proportion of National e-Infrastructure Survey respondents indicated that they had training materials that they would be happy to share, and that they would be keen to collaborate on shared training activities. A prototype "one stop shop" for an e-Infrastructure training website has been developed, which is available at http://etraining.esc.rl.ac.uk/.

With the launch of ARCHER, the UK's new national supercomputing facility, the opportunity had been taken to move to a more flexible delivery model for training courses, with a larger number of slots at locations around the UK.

An e-Infrastructure "driving test" had been developed by EPCC and the SSI in collaboration with DiRAC - the integrated supercomputing facility for theoretical modelling and HPC-based research in particle physics, astronomy and cosmology.

This "driving test" functions as Level 0 in the proposed DiRAC Training Framework, listed below, that will be established during 2014:

Level 0 – Can I use a system?
Level 1 – How do I use a system?
Level 2 – What can I do on the system?
e.g. basic scripting, parameter sweeping
Level 3 – Advanced courses e.g. ARCHER, EBI
Level 4 – Practitioner, advanced methods
Level 5 – Expert

1.6 Funding for Code Development and Maintenance

The 2013 National e-Infrastructure Survey report recommended that an investment be made in strategically important codes. Following on from this we have seen renewed funding for Collaborative Computational Projects (CCPs), the EPSRC Software for the Future initiative, and funding for industrial exploitation of "big data" through InnovateUK's (TSB) recent Data Exploration call.

[2] http://software.ac.uk/attach/ReportOnWorkshopForEInfrastructureTrainers.pdf

In addition InnovateUK (TSB) ran a £10M funding call in the area of Data Exploration (March 2014) with BBSRC and EPSRC.

Later in 2014 and early 2015 there will be opportunities for software, software developers and code development under the Horizon 2020 programme.

From the Software for the Future call[3], current at the time of writing:

> In both the recently published EPSRC e-infrastructure roadmap and the EPSRC Software as an Infrastructure strategy, the importance of software development and the need to invest in people and training in this area has been strongly highlighted. EPSRC has therefore made a long-term commitment to support software development, ensuring that funding continues to support leading scientific research and key codes used by the Engineering and Physical Sciences community.
>
> Subject to quality, up to £4M of funding is available for projects focused on the development of software that is used in computational science and engineering. All proposals submitted to this call must fall within the EPSRC remit.

[3] http://www.epsrc.ac.uk/funding/calls/2014/Pages/softwarefuture.aspx

2. Capital Investments in Big Data Facilities

One of the main activities over the last year has been to set up Big Data projects using funds announced by the Government in December 2012. Major Awards have been made to 18 centres in the UK, 16 of whom are HEIs. 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, which are listed in the Table below:

Big Data Project	**RC**	**Amount (£M)**
Energy Efficient Computing for the Square Kilometre Array (£11M) and Industrial Computing (Hartree Centre. £19M)	STFC	30
Digital Transformations in Arts and Humanities:	AHRC	8
e-infrastructure for Biosciences	BBSRC	13
Research Data Facility and Software Development	EPSRC	8
Administrative Data Centres	ESRC	36
Understanding Populations	ESRC	12
Business Datasafe	ESRC	14
Biomedical Informatics	MRC	55
NERC Environmental Big Data (http://www.nerc.ac.uk/funding/available/nationalcapability/envinfo/)	NERC	13
Total		**189**

The Medical Research Council (MRC) is investing £50 million in bioinformatics for health, medical and administrative data:

- £20M for the four Farr Institute nodes, for e-Infrastructure and buildings, June 2013, in addition to the initial £19M recurrent 5yr funding from MRC and 10 other funders to establish the four centres;
- MRC £39M for six Medical Bioinformatics Initiative projects, Feb 2014:
    - Building new ways of linking across complex biological data and health records to solve key medical challenges;
    - UCL-QMUL-LSHTM-Crick-Sanger-EBI-KCL; Data-driven discovery for personalised medicine;
    - Oxford; Oxford Big Data Institute;
    - Leeds; Leeds MRC Medical Bioinformatics Centre;
    - UVRI Uganda Research Institute, Sanger; Medical Informatics Centre, Cambridge;
    - Warwick and Swansea; Medical Microbial Bioinformatics;
    - Imperial; Medical Bioinformatics partnership.

- 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 £9 million in 4 new Administrative Data Research Centres (ADRCs) at Edinburgh, Swansea, Belfast and Southampton, with Retail DRCs at UCL and Leeds University, 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.
    - Phase 1 of the ESRC funding will enable researcher access to non-health government data such as education, benefits, and tax records. The Administrative Data Research Centres are at Dundee (ADRC-Scotland), Swansea (ADRC-Wales), Belfast (ADRC-N Ireland) and Southampton-UCL-IoE-IFS (ADRC-England), with the Administrative Data Service at Exeter, with all forming the Administration Data Research Network;
    - Phase 2 of ESRC funding includes the Retail datasafe centres at UCL and Leeds;
    - Phase 3 is on hold while ESRC re-scopes the call.

- 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, which is designed to provide research data management and analysis services for ARCHER and potentially other Ne-I projects.

Other data infrastructure investments include the European Bioinformatics Institute at Hinxton (Cambridge) which received £75M from BBSRC and the Open Data Institute, a not for-profit institute funded by Industry and InnovateUK (TSB) (£10M over 5 years, subject to industry investment). The Open Data Institute is a big data undertaking and is 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 has earmarked £100M to sequence 100,000 genomes in rare diseases, cancer and infectious diseases, delivered through NHS England by Genomics England Ltd. Genomics England is working with InnovateUK (TSB) via the Small Business Research Initiative to help develop technology to understand the genetic

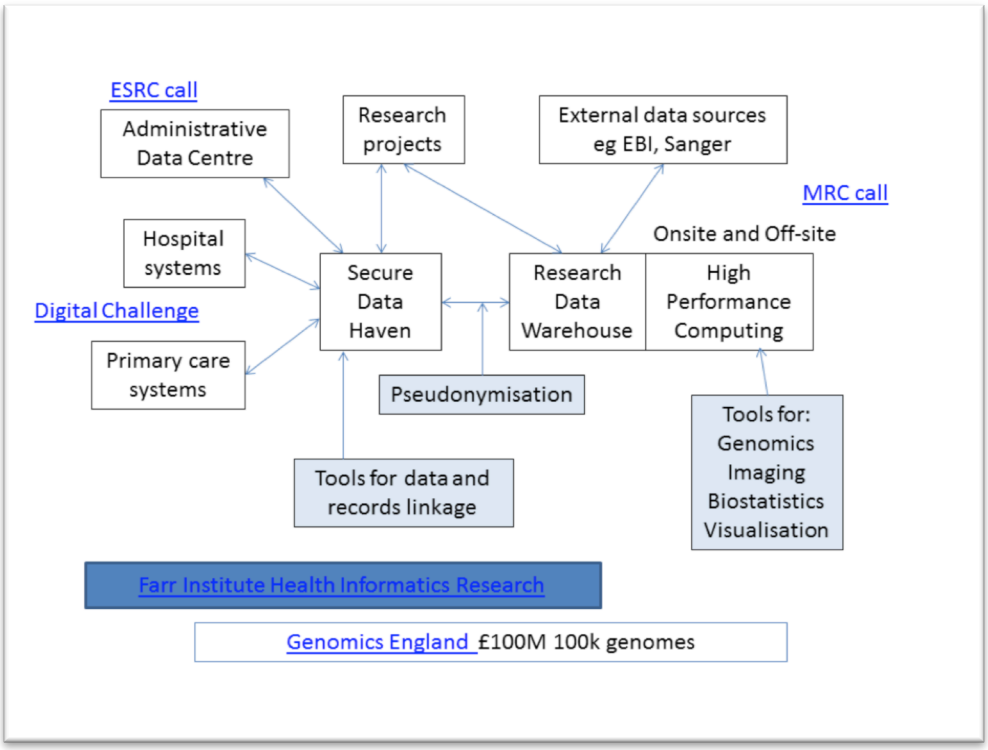
outputs from the DNA sequences, total £18M, with first £10M already announced[4].

£11M was awarded to fund energy efficient computing for the SKA Project and £19M to set up an energy efficient computing infrastructure at the Hartree Centre.

2.1 Particular Issues around Access

In the so-called *people* areas of ESRC and MRC and Health related research, there is an added issue over data security and secure access to data. The figures below (provided by Dr. Jacky Pallas, UCL) show an indicative example of possible data flows between various data sources. The Secure Data Haven model will allow these data to be analysed, post-processed, linked and explored in a secure environment.

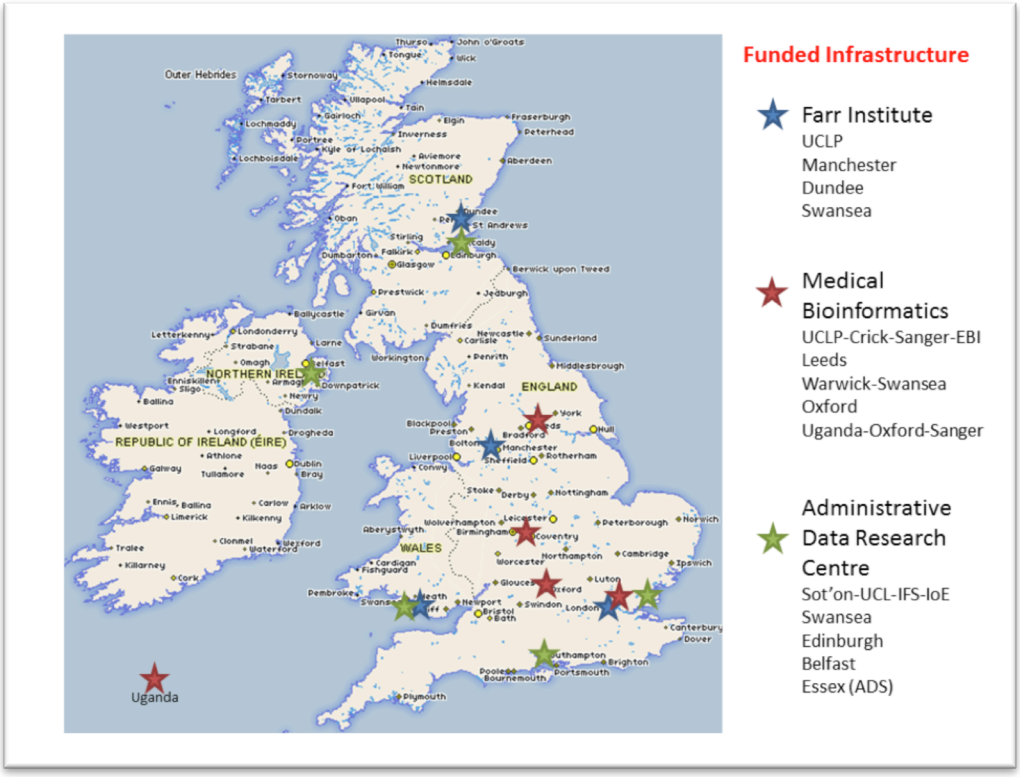
The figure includes an example of a psuedonymization layer, between the Secure Data Haven and Research Data Warehouse. As this work progresses the nature and strength of pseudonymization will become clear, and also the boundaries and interfaces between systems where pseudonymization, true anonymization and enhanced information assurance will be required.



The next figure below shows the location of the awards to the Farr Institutes, Medical Bioinformatics and Administrative Data Research Centres. These have been made to 18 centres throughout the UK. The projects have highlighted requirements for secure, but common AAAI, to enable effective single sign on within research domains and if

[4] https://www.gov.uk/government/news/mapping-rare-disease-and-cancer-genes-competition-winners

necessary between domains. This has led to an ongoing collaboration between the three funded projects and JANET to provide the necessary AAAI to make access and linkage of datasets easier and more straightforward than hitherto. In addition the need for secure data transport has also been highlighted by these projects.



3. Who are the Ne-I's Service Providers

3.1 HEIs

The 2014 Survey shows that 25 HEIs are providing services to National e-Infrastructure Projects.

HEIs provide High Performance Computing and High Throughput Computing Facilities as follows (the number of participating HEIs is in parentheses): the National Service (1); DiRAC (4); GRIDPP (13) and the six Regional HPC Centres (13)

16 HEIs are providing services to Big Data Projects such as the GRIDPP (13), JASMIN (3), SKA (1), Research Data Facility (1); Farr Institutes (7), eMEDLAB (7) and the Administrative Data Centres (6).

Most of these HEIs are providing more than one service to the National e-Infrastructure.

These 25 HEIs receive the majority of RCUK research grants.

HEIs are playing a key role in delivering advanced IT services. Their role is mission critical to the National e-Infrastructure as they manage most of the resources in the National e-Infrastructure.

The contribution of HEIs can be summarised as follows:

- Innovation; this comes from their diverse research agendas that operate on local, national and international scales;
- HEIs are agile and nimble in creating services that respond to the research agenda. The 3 year timescale of the research grant or post-graduate degree research is the driver here;
- HEIs dynamically create the Ne-I roadmap from both their research agendas and their immediate needs to respond to that changing research agenda;
- This ensemble of activities produces a vibrant and diverse set of services, which are effective for a wide range of activities;
- HEIs are used to collaborating and setting up both short-lived and long-term structures to deliver common research goals.

The Issues:

- Funding is irregular and piecemeal:
    - The mission critical contribution of HEIs to the Ne-I does mean that HEIs do need access to Core Baseline Funding to maintain their local e-Infrastructure;
    - **Such access to funding would give an opportunity for HEIs to consider matching their own local e-Infrastructure funding to Ne-I funds to upgrade local networks, generate new activities, especially in the area of software support, training and industrial engagement. HEIs could develop business plans that maintain the links between**

research objectives and funding, create a local critical mass of computational scientists and software engineers and give HEIs an opportunity to grow a supporting income for their e-I activities;

- Staffing is at levels that simply keep the machines running;
- Most Ne-I users work in HEIs:
  - **It is necessary to build up software expertise at the practitioner and expert level using locally organised training;**
  - **HEIs need to seed-corn the establishment of research software engineer groups for use by their researchers, which includes career paths for these software engineers;**
  - **HEIs needs to create academic posts that teach research software engineering in the same way that other engineering disciplines are taught and researched**;
  - **A study should be undertaken to identify key outputs and innovations that would not have happened without this investment.**

3.2 National and International "Centres"

If HEIs serve most Ne-I users and provide the local face of the Ne-I, then what is the role of our National Centres?

Should they have key underpinning roles and long term development tasks that allow the Ne-I to function? JANET clearly does this for Networking Services.

The National Centres support those research projects that are too big and complex to be served by HEIs.

The National Centres also play "hub" roles within several national projects in partnership with HEIs.

The National Centres also contain the critical mass needed to retain and develop key skills, expertise and domain specific knowledge for the Ne-I.

We present below a tentative list of National Centres for the UK National e-Infrastructure:

- JANET (JISC) provides key Networking Services to the UK academia. The new SJ6 Backbone will provide new capabilities in terms of data rates, but also data assurance. In addition JANET (JISC) are leading developments in Access and Information Security and access to Public Clouds;
- Edinburgh Parallel Computing Centre (EPCC), which is based at the University of Edinburgh, provides programming, hosting and operations services for the National Service, including ARCHER and DiRAC (the BlueGene/Q). EPCC also offers computing, software and project related services to academia and industry;

- The MRC Francis Crick Institute will play a leading role in providing and organising compute and data resources for the UK's Biomedicine communities. In addition the Wellcome Institute and the Institute of Cancer Research provide services via a network of centres in the UK;
- STFC's Scientific Computing Division (SCD) provides compute, data and programming services to National e-Infrastructure Projects both solely and in partnership with HEIs; examples of this are GRIDPP and JASMIN/CEMS. In addition they provide computational and data services to large Facilities such as Diamond;
- STFC Hartree Centre provides advanced programming services to UK Industry. It also hosts an energy efficient computing development centre;
- SKA has recently opened an Open Architecture Laboratory in Advanced IT to drive the IT developments needed to provide future SKA requirements;
- The EBI, The Sanger Centre and The Genomic Analysis Centre provide computational and data services to the UK's Life Sciences communities;
- **There is now the opportunity to formally recognise and define the contribution of the National "Centres" to the Ne-I and to make sure they are adequately funded for carry out their individual missions.**

4. The National E-Infrastructure and Industry

The Survey does show the providers who have long and close relationships with industry and commerce: e.g. EPCC, the Hartree Centre, HPC Wales, The Cambridge HPCS service (SES5), Loughborough (HPC Midlands) and Warwick (Midplus). Newer services such as N8 and DiRAC are now starting to attract funding and interest from Industry, but this is still a very immature activity in these services.

Experience shows that Business Developer Officers are key to ensuring long term success, by both creating and developing the relationships necessary to build up a sustainable activity. Very few providers have such Officers and their current staffing is simply directed to keeping the systems running. EPSRC recognised this was an issue for the Regional HPC providers and granted seed-corn funding to 3 providers (£50k each) for SME engagement in late 2013.

The Project Directors Group, which now includes InnovateUK (TSB) and the KTNs, has focussed on dealing with infrastructure issues that prevent access to our services. These include:

- JANET have made it easier for non-academic customers to access their network, subject to contract and charges. Ne-I members have helped JANET test their processes;
- The work on providing a secure single sign on for UK academia (led by JANET) is also just as useful for Industrial users and should provide a single login for industrial users as well as academics;
- JISC plan to provide portals that allow users to submit jobs to Ne-I resources. These can be for specific applications or be of a more general nature;
- The RCUK Security and Access Group (led by JANET) will give advice and help to providers who wish to get the appropriate accreditation for work within particular sectors. A “matrix of requirements and accreditation” will be produced as not all problems require the same level of information assurance;
- InnovateUK (TSB) and JISC are producing an application to allow prospective industrial users to find out where resources are located and to give a description of them, including contact details. The 2014 Survey data will be included in this application;
- InnovateUK (TSB) and DiRAC plan to produce a Prototype Skills Database that can be accessed by Industry and Academia alike. This would allow easier access to our broad computational science skill sets, knowhow and research experience.

These activities should be concluded and made available to the relevant communities.

It is being proposed that access to relevant HPC and Big Data Services should be part of the function of so-called on-ramp centres that could be housed in Catapult Centres. These would mainly focus on helping SMEs to make use of the Ne-I’s HPC and Big Data Services.

The main services would be:-

- Outreach to the individual sectors to make SMEs aware of the HPC and Big Data services available;
- Link potential customers with service providers;
- Examples of services provided would be:
    - Access to CPU and data services to run own applications;
    - Access to commercial software applications to run on CPU and data services;
    - Access to consultancy and software engineering services.
- Provide access to portals, information, and training on how to access these services.

Access to Commercial Application Software remains a real show-stopper. Anecdotal evidence from a few Ne-I projects now shows that this is preventing HEIs, for instance, from even starting some projects as the start-up costs are simply too high.

The work that is being overseen by the PDG and the RCUK Ne-I Group should support the operation of the proposed on-ramp centres.

This area was difficult to survey as providers are not required to say how much time was used for commercial use. The survey question requested the peak use of the system by industry and commerce.

There was a poor response to this question, suggesting that well over the 50% of Ne-I providers have little contact with non-academic customers.

The decision by InnovateUK (TSB) to set up and support an e-Infrastructure Special Interest Group to drive more interaction between e-Infrastructure providers and industry and commerce is therefore vindicated and is a welcome addition to activities co-ordinating and focussing on this vital area.

The InnovateUK (TSB) e-Infrastructure Special Interest Group should survey the Ne-I as a matter of some urgency to get a realistic understanding of the level of commercial activity in the academic research part of the Ne-I.

5. Conclusions and Recommendations

We have seen significant investment in the core of the JANET network and in e-Infrastructure facilities connected to it. Campus network infrastructure is a critical part of the institution's capacity to exploit e-Infrastructure investments. **Internal investment by institutions is required to ensure that this remains fit for purpose.**

Historically (up to 2005) HPC focussed e-Infrastructure services have tended to work on the basis that their users are predominantly writing or working with parallelised codes. However it is increasingly the case that the majority of users just wish to run much simpler (though sometimes embarrassingly parallel) codes on a "bigger" machine – with more RAM per core, storage, and so on. Cloud computing technologies such as OpenStack and Linux Containers / Docker can potentially play a major part here, with a standard library of virtual machine appliances preconfigured for common scientific computing workflows. **JASMIN, Sanger Institute, GRIDPP, SKA and DiRAC should plan to work together to explore the practicalities of this approach and show that resources within Research Domains can be configured into effective and efficient private clouds that allow researchers to run their workflows easily on a domain private cloud or on resources in another part of the Ne-I.**

There is a common perception that commodity cloud computing services are not cost effective for compute and data intensive applications. Through the recent JANET cloud computing service agreements to set up portals to Public Cloud Providers, good relations have been built up with the major service providers and it is now a good time to expand the range of service providers available to the National e-Infrastructure. **JANET will take this work forward, in collaboration with major e-Infrastructure service users and commodity cloud providers. However the costs and benefits of such access will have to be carefully measured and it is unlikely to be a solution for many of our problem sets in the next 2-3 years. It is clear that the Public Cloud Providers and Ne-I providers will have to exchange information concerning the problem types and sizes in the Ne-I, so that expectations are not unduly raised and that methodologies are built up that allow effective and economic use of Public Clouds.**

There has been widespread interest in the Moonshot technology trial, but this in itself only addresses part of the problem of reducing the friction of access to e-Infrastructure services. It is necessary to build infrastructure around communities and domains. Moonshot can provide the existing infrastructure for identity authentication management but not the bits to do the federated access control that is tailored to the needs of individual domains and/or communities.

There are a number of further issues around authorisation and accounting, information assurance and the legal and contractual framework within which e-Infrastructure services are shared between organisations – potentially leading to a single e-Infrastructure service portal for each research domain and/or research community. **JISC is well placed to continue coordinating efforts with the Research Domains and Innovate UK (TSB) in this area, building upon existing work wherever possible.**

Whilst concerted programmes of work exist around large and specialist facilities, e-Infrastructure provision in HEIs have not had the same level of support or incentives to enhance investments in this key area. It is clear from the survey results that many institutions are unable to reap the full benefits from their e-Infrastructure investments. This is due to the overhead involved in "keeping the show on the road". Survey respondents tell us that this often comes close to (or exceeds) the available FTE count. We recommend that the funders consider potential incentives to improve this situation. **The e-Infrastructure Leadership Group is encouraged to consider potential approaches such as greater regional collaboration, supported by an element of matched funding for HEIs' e-Infrastructure investments.**

There is a clear increase in the awareness of the need for support for Software, Software Developers and Training. The clear standouts were *career paths for developers of scientific software* and *basic software engineering training for our user-researchers*. For software issues, scientific consortia highlighted optimisation methods, recognition of software as a research object / credit for software and recognition of software planning / sustainability in grants by investigators and reviewers. Institutions highlighted, reproducibility / correctness of results, sustainability of software, more robust Linux installation processes, general training for staff and postgraduates, licensing, software as a research output and new technology expertise.

For training, scientific consortia felt that the following topics were underprovided: Basic software engineering skills; using clusters / parallel and distributed programming; Data analysis techniques; choosing and using software appropriately and Bid writing, which includes software. Institutions highlighted: basic software engineering skills; How to apply computational techniques as research tools; data analysis, data curation and management; courses tailored for X-informatics subjects and parallel programming.

It was felt that training is still undervalued as an activity and is expected to be provided for free. There needs to be a clear Branscomb pyramid in terms of Training and Support.

Funders should make applicants aware that it is permissible to apply for research software engineer time on grants, and that it is appropriate to class these as research staff on grants where the work to be carried out involves a significant research and development aspect.

The work to raise the profile of the research software engineers should continue; in particular the role, value and potential career paths should be highlighted in submissions to the ELC and HEIs. The value of dedicated developer support at HEIs should also be highlighted to Pro-VCs for research and Directors of Research, as well as successful funding models.

Training materials and courses should be easily available. A service to host materials, advertise courses and assign courses to particular levels of skill and knowledge would be very helpful in making sure users get appropriate training and are able to progress easily from level to level. HEIs and Research Domains

should be encouraged to share what materials and resources they can with each other. Training has to be provided for trainers.

Appendix A: List of Surveyors and Acknowledgements

Jeremy Yates, Harpreet Dhanoa and Clare Jenner (DiRAC), Martin Hamilton (JISC), Neil Chue Hong and Simon Hettrick (Software Sustainability Institute) and Oz Parchment (Southampton).

Further material provided by Andrew Richards (Oxford) and Alan Real (Leeds).

Many thanks to all those who provided substantive contributions to this report, including Jacky Pallas (UCL), Andrew Stewart (Bristol), Lydia Heck (Durham), Dugan Witherick (UCL,Oxford), James Hetherington (UCL), Mark Parsons (EPCC) and Susan Morrell (EPSRC).

Appendix B: Who Will Receive this Survey

The National e-Infrastructure Project Directors Group, the RCUK National E-Infrastructure Group and the BIS e-Infrastructure Leadership Council.

Appendix C: List of Respondents

Large and Specialist Services:

- National HPC Service: ARCHER and Research Data Facility (EPSRC and NERC);
- The (MRC) Francis Crick Institute;
- Sanger Institute (Wellcome);
- Institute of Cancer Research;
- National Oceanography Centre;
- Regional HPC Services (HPC Wales, N8, HPC Midlands, ARCHIE-WeSt);
- The Genome Analysis Centre (BBSRC);
- STFC Scientific Computing Department:
    - Data and Systems Divisions: HPC for STFC experimental facilities and GRIDPP data for the Large Hadron Collider;
    - Hartree Centre: HPC and Big Data Applications and Advanced Software and Services for Industry and Industry-Academia Collaboration;
    - Applications Division: Support of HPC applications and the CCPs;
- DIRAC – HPC and Big Data for Theoretical Particle Physics, Astrophysics, Cosmology and Nuclear Physics (STFC);
- GRIDPP – HPC for UK Large Hadron Collider Science (STFC);
- Wellcome Trust Centre for Human Genetics (Oxford).

Service Management Responses Only:

- The Farr Institutes of Health Informatics Research (MRC);
- Administrative Data Research Centres (ESRC);
- Administrative Data Service (ESRC, Essex);
- JASMIN and CEMS services (NERC);
- Regional HPC Services Midplus;
- AHRC – no inputs;
- EMBL European Bioinfomatics Institute (BBSRC).

JANET, The Met Office, and ECMWF were not surveyed as they make their own reports to the ELC and RCUK.

HEIs:

- Cardiff University;
- DAMTP, University of Cambridge;
- Durham University;
- Imperial College London;
- King's College London;
- Lancaster University;
- Loughborough University;
- QMUL;
- Queens University of Belfast;
- The University of Sheffield;
- UCL;
- UCL Computer Science;
- University of Aberdeen;
- University of Bath;
- University of Birmingham;
- University of Bristol;
- University of Central Lancashire;
- University of East Anglia;
- University of Edinburgh;
- University of Essex;
- University of Glasgow;
- University of Huddersfield;
- University of Leeds;
- University of Leicester;
- University of Liverpool;
- University of Manchester;
- University of Oxford;
- University of Southampton;
- University of Strathclyde;
- University of Sussex;
- University of Warwick.

Appendix D – Service Management: The Survey Questions

UK National e-Infrastructure Inventory 2014 – Service Management

Q1. Organisation name?
Q2. Organisation Unit?
Q3. Respondent's email address?
Q4. Respondent's job title?

Budget

Q5. Is your HPC Budget ring fenced, or do you have to make a fresh case for support each time?
Q6. Is your primary budget for HPC CAPEX or OPEX?
Q7. Are you directly charged for power and cooling or other Estates costs?
Q8. If your answer to the previous question is yes, do you capitalize Estates costs, or pay then from your OPEX budget?
Q9. Do you outsource or plan to outsource any part of your HPC provision?
Q10. Which areas of HPC provision are you (or might you) outsource?

Staffing

Q11. How many FTE support your HPC activity?
Q12. FTE count if >4?
Q13. Where are the staff involved in supporting HPC based?
Q14. Breakdown of effort by %FTE?

Training and User Support

Q15. Do you offer your own training courses?
Q16. URL of training courses website if applicable.
Q17. Training materials – do you produce your own?
Q18. URL of training materials website if applicable.
Q19. Are you interested in talking to SSI about contributing to a shared pool of training and/or training resources?
Q20. What proportion of your researchers are in self-supporting research groups?
Q21. What sort of help does a new person typically need to use your systems?
Q22. Other forms of support?
Q23. Are you able to provide HPC case studies for (e.g.) HPC-SIG website, RCUK and InnovateUK (TSB)?

Further Information

Q24. Do you have a Research Management Policy?
Q25. URL of Research Data Management Policy if applicable?
Q26. What technologies have you deployed that can be regarded as being for Data

Exploration (“Big Data”) activities?
Q27. Further Information.

The survey results are open data, licensed under the Creative Commons CC0 (CC Zero) license, which permits re-use without attribution.

https://docs.google.com/a/lboro.ac.uk/spreadsheet/ccc?key=0Auw-AE5DOPfNdHpNRnhHVzVrb0Y2a0FMVVhoa0hueEE&usp=drive_web#gid=0

Appendix E – Service Management: Summary of the Survey Data

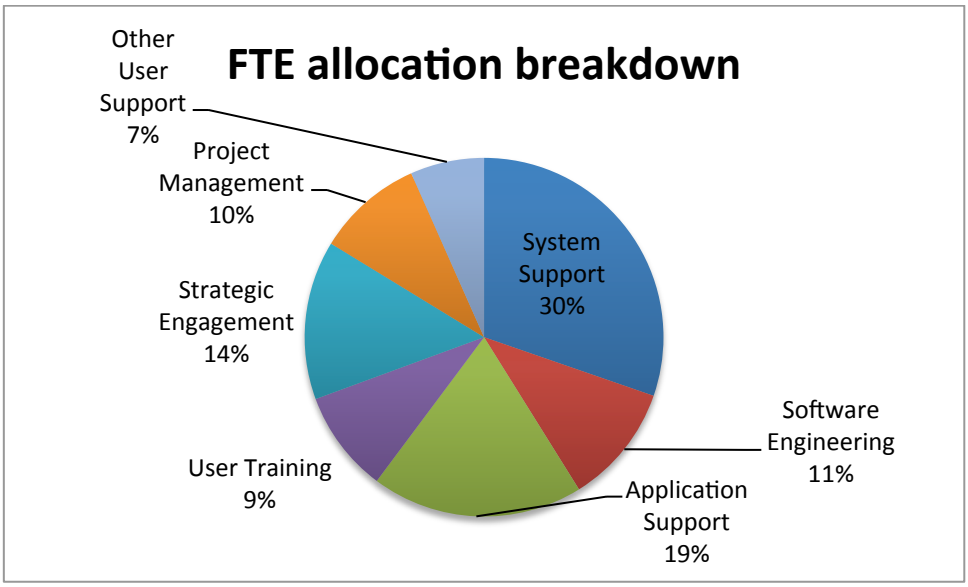
- We received 17 responses from Large and Specialist Projects;
- We received 30 responses from the 38 HEIs that are members of the HPC-SIG – See Appendix C for a full list of respondents;
- A summary of the responses is given below and any further information provided by respondents.

Funding

- 54% of HEI respondents had to make a fresh case for support to their organization for each HPC refresh cycle, with 39% having an either wholly or partly ring-fenced budget element for HPC;
- Some HEIs were charging back HPC time to service users according to the Small Research Facility model;
- 74% of HEI respondents indicated that their HPC budget was wholly capital expenditure (CAPEX) based, with 15% funding HPC from operating expenditure (OPEX), and 11% a mixture of both;
- 77% of all survey respondents reported that they were not directly charged for power and cooling and other Estates costs. This figure breaks down to 85% of HEIs not being directly charged, and 63% of large and specialist facilities;
- 73% of the respondents that had indicated that they were charged for Estates costs, were funding them exclusively from operating expenditure.

Service Provision

- 14% of all respondents presently outsourced some aspects of HPC provision. These figures were consistent across large/specialist facilities and HEIs;
- Current or projected use of outsourcing was mainly around strategic aspects such as national Capability systems and "cloud-bursting" when in-house resources were unavailable or fully committed;
- Large/specialist facilities had a median FTE count of 3, standard deviation of 14 and a modal value of two. This reflects the split between the major HPC employers (Hartree, EPCC, HPC Wales, and STFC) and smaller and/or more specialized institutions such as TGAC and the Sanger Institute;
- HEIs had a total headcount of 85.4, median FTE count of 2.5, standard deviation of 2.3, and a modal value of 2. Here there are a small number of outliers – such as Research Information Services at UCL, which employs staff in a range of roles including research facilitators, and research software engineers;
- 67% of HEIs and 50% of large/specialist respondents were operating a centralised support model for HPC. 19% of HEIs and 46% of large/specialist research organizations were operating a distributed model with both central and localised support. 15% of HEI respondents ran HPC facilities within their department or research centre.



Training, Documentation and User Support

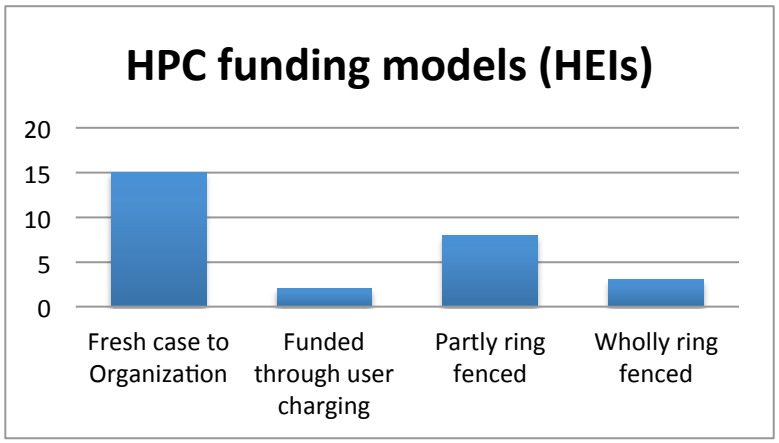
- 72% of respondents overall offered training courses, with several respondents also noting that they promoted national level (e.g. ARCHER) training courses to their service users;
- 85% of all respondents indicated that they produced their own training materials, although several respondents noted that they were exploring ways of sharing training documentation or customizing existing course materials;
- 72% of all respondents were interested in engaging with the Software Sustainability Institute to develop a shared pool of training and/or training resources. 4% were already engaged with this process;
- Both large and specialist services and HEIs reported that a high proportion of their users were working in self-supporting research groups, however 36% of respondents overall had a significant number of users that were not part of a larger group;
- Linux orientation was by far the most common support requirement for new users, with 35% of all respondents noting this as a key area. 11% of respondents stated that new users often required an introduction to programming before they could make use of the service;
- Respondents were asked about other forms of user support offered as part of their HPC service provision. Several noted that they promoted courses available through the likes of EPCC, NAG, the European Grid Initiative and WLCG (Worldwide LHC Computing Grid), or through Doctoral Training Centre provision;
- 28 respondents had case study material they would be willing to contribute to HPC-SIG, RCUK, InnovateUK (TSB) etc;
- 27 respondents had a Research Data Management policy in place (for URLs see Appendix F below);
- Where institutions were involved in broader community-wide service provision it was not always clear when a respondent was speaking from an institutional perspective. This was perhaps due in part to the accelerated timescales for the survey;
- Some respondents regarded their organization as data intensive but not a user of “HPC” per se. Future iterations of the survey will aim to address this.

Appendix F – Service Management: Full Break Down of Survey Data

1. Is your HPC budget ring fenced, or do you have to make a fresh case for support each time?

The funding situation for large and specialist projects is well understood, but the situation around institutional HPC has been less clear. Responses from 28 HEIs indicated that it was unusual to have a wholly ring fenced funding model in this area, with 54% of respondents having to make a fresh case to their organization each time.

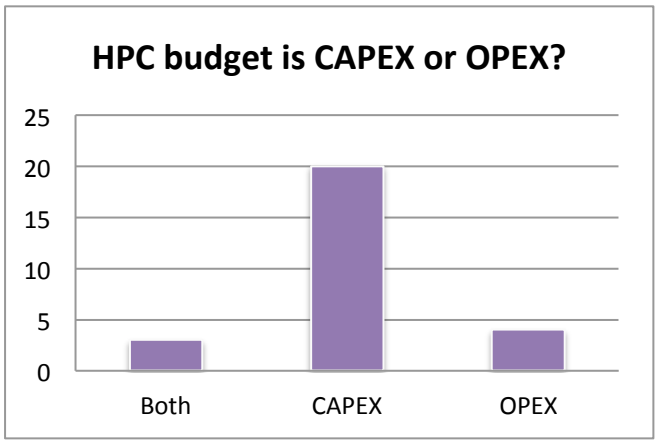
HPC funding models		
Fresh case to Organization	15	54%
Funded through user charging	2	7%
Partly ring fenced	8	29%
Wholly ring fenced	3	11%



2. Is your primary budget for HPC CAPEX or OPEX?

74% of HEI respondents indicated that their HPC budget was wholly capital expenditure (CAPEX) based, with 15% funding HPC from operating expenditure and 11% a mixture of both.

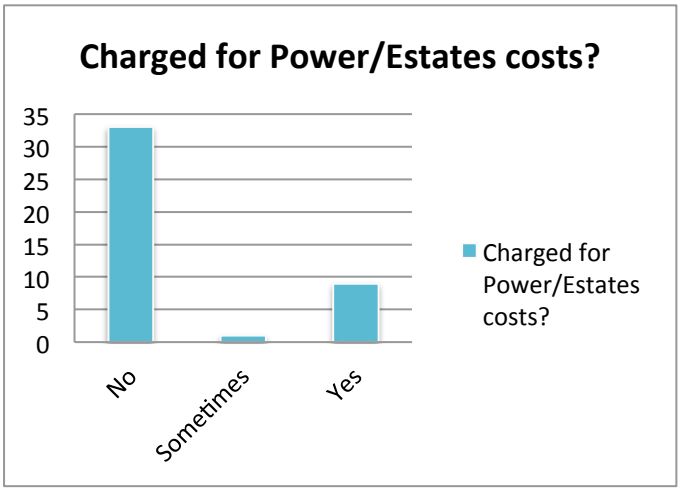
HPC budget is CAPEX or OPEX?		
Both	3	11%
CAPEX	20	74%
OPEX	4	15%



3. Are you directly charged for power and cooling or other Estates costs?

77% of all survey respondents reported that they were not directly charged for these costs. This figure breaks down to 85% of HEIs not being directly charged, and 63% of large and specialist facilities.

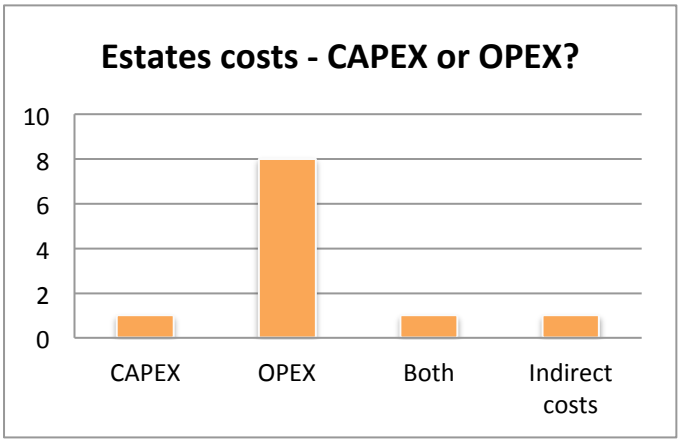
Charged for Power/Estates costs?		
No	33	77%
Sometimes	1	2%
Yes	9	21%



4. If your answer to the previous question is yes, do you capitalise Estates costs, or pay them from your OPEX budget?

73% of the respondents that had indicated that they were charged for Estates costs were funding them exclusively from operating expenditure.

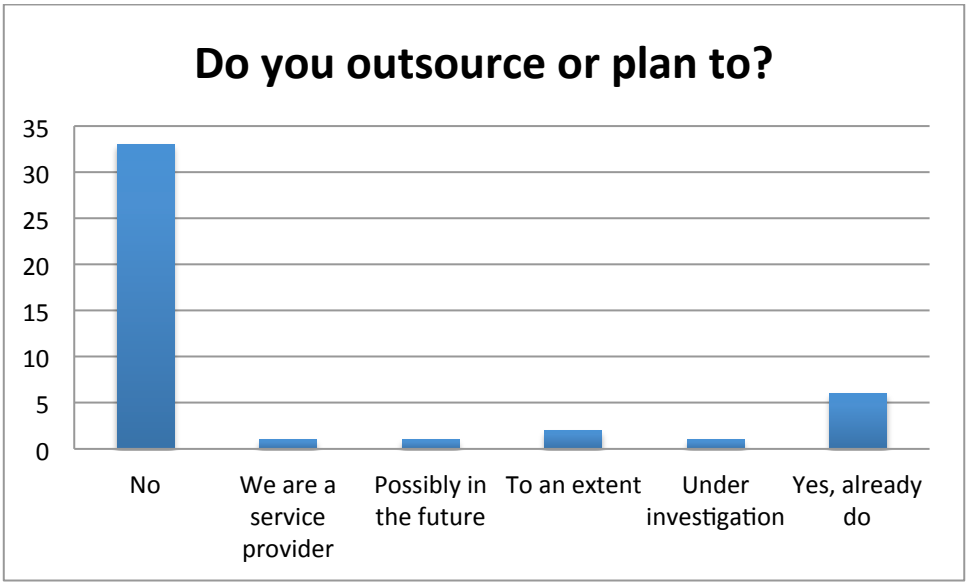
Estates costs – CAPEX or OPEX?		
CAPEX	1	9%
OPEX	8	73%
Both	1	9%
Indirect costs	1	9%



5. Do you outsource or plan to outsource any part of your HPC provision?

75% of all respondents did not presently outsource any parts of their HPC service provision and had no plans to. 14% of all respondents presently outsourced some aspects of HPC provision. These figures were consistent across the subset of HEIs surveyed.

Do you outsource or plan to?		
No	33	75%
We are a service provider	1	2%
Possibly in the future	1	2%
To an extent	2	5%
Under investigation	1	2%
Yes, already do	6	14%



6. Which areas of HPC provision are you (or might you) outsource?

Respondents indicated that they are presently (or would consider) outsourcing the following aspects of HPC service provision:

- Training;
- End-to-end service provision / system hosting / service management;
- Large scale “capability” requirements;
- Burst capacity when in-house resources committed or unavailable;
- Work requiring access to national Capability resources or regional collaboration.

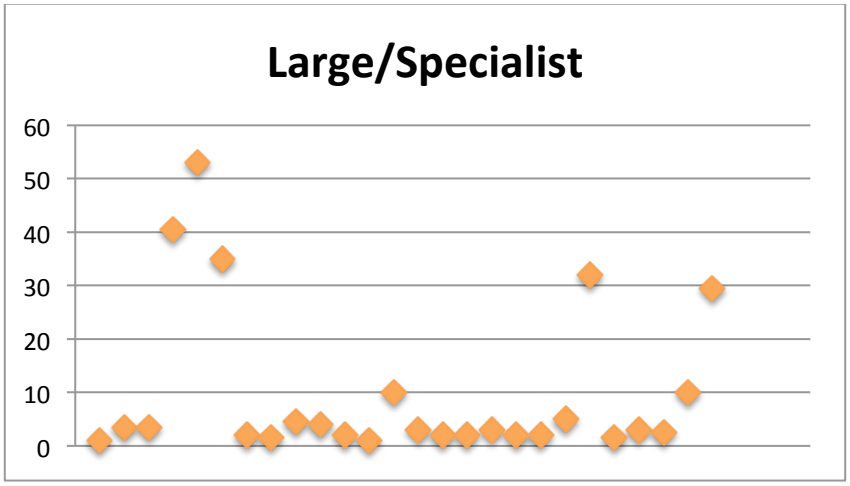
Respondents also noted that they had considered adopting a wholly cloud based model for HPC service provision, but discounted this due to CAPEX funding model and heavy usage requirements leading to a poor return on investment.

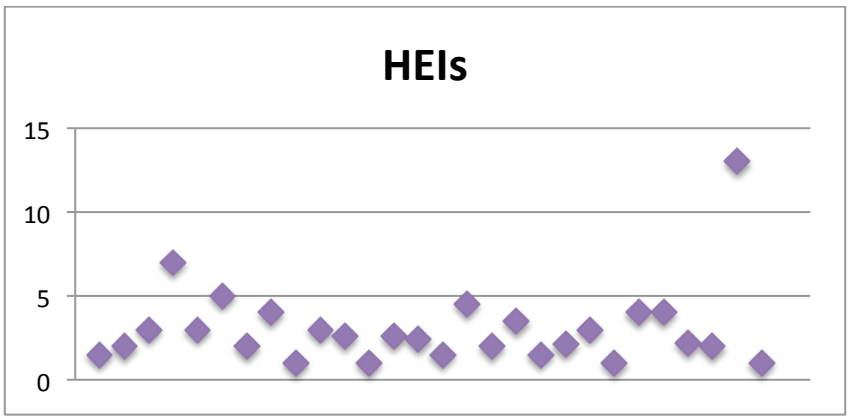
7. How many FTE support your HPC activity?

Please see answer to Q8 below.

8. FTE count if >4

Large/specialist facilities had a total headcount of 259, median FTE count of 3, standard deviation of 14 and a modal value of two. This reflects the split between the major HPC employers (Hartree, EPCC, HPC Wales, and STFC) and smaller and/or more specialized institutions such as TGAC and the Sanger Institute.



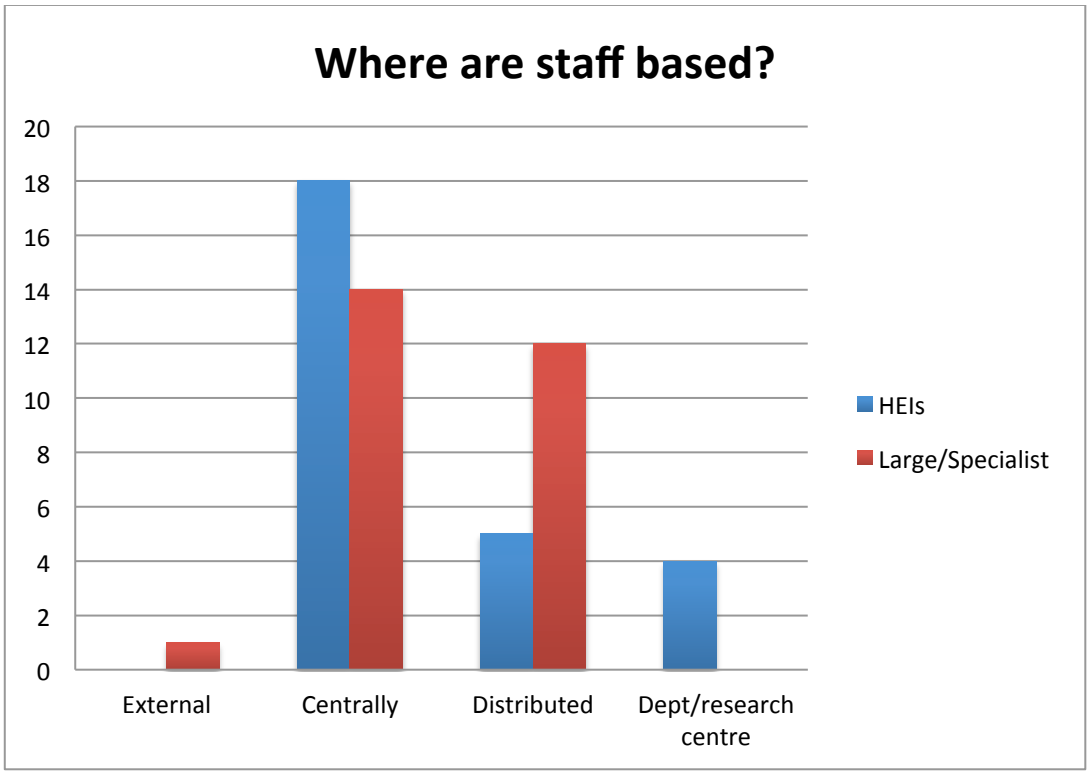


HEIs had a total headcount of 85.4, median FTE count of 2.5, standard deviation of 2.3 and a modal value of 2. Here there is one clear outlier – Research Information Services at UCL, which employs 13 FTE in a range of roles including research facilitators and research software developers.

9. Where are staff involved in supporting HPC based?

67% of HEIs and 52% of large/specialist respondents were operating a centralised support model for HPC. 19% of HEIs and 44% of large/specialist research organizations were

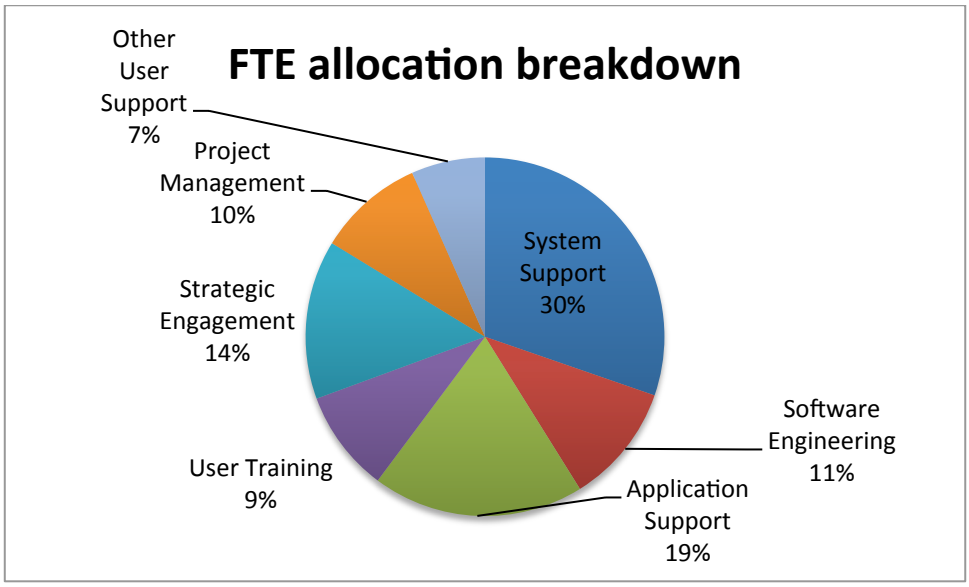
operating a distributed model with both central and localised support. 15% of HEI respondents ran HPC facilities within their department or research centre.



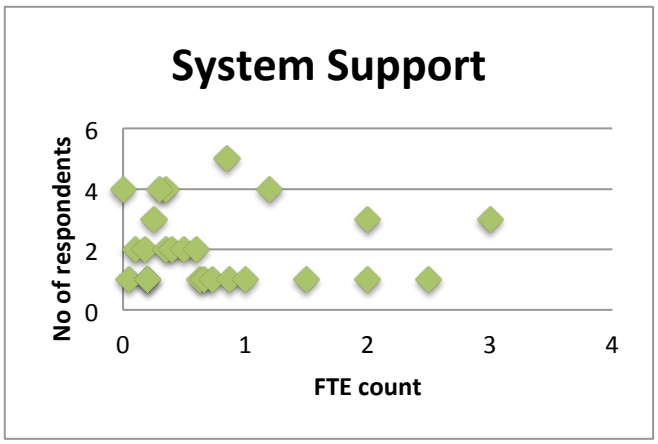
10. Breakdown of effort by %FTE

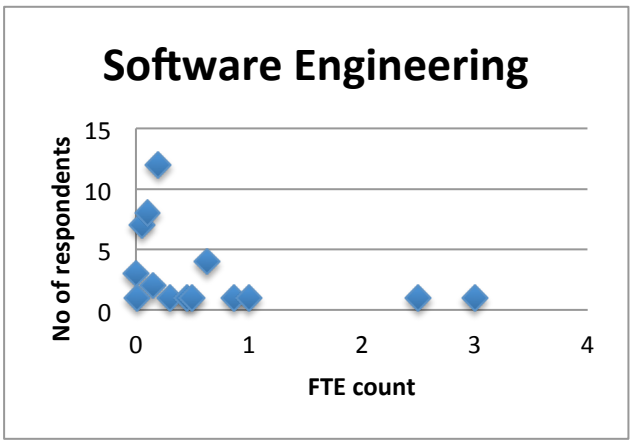
Respondents were asked to break down the FTE effort allocated to particular tasks in relation to HPC service provision. Please note that these figures include both large scale/specialist centres and provision at HEIs.

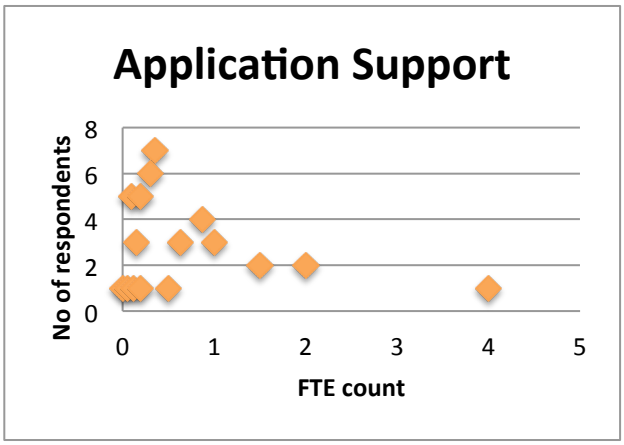
At first glance there appears to be a significant skew in HPC support staff FTE time allocation towards system support, which may be a reflection of the low staffing levels in this area at many HEIs and specialist research institutes.

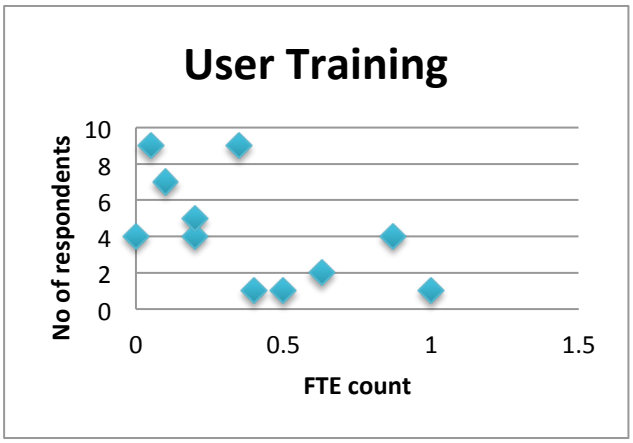


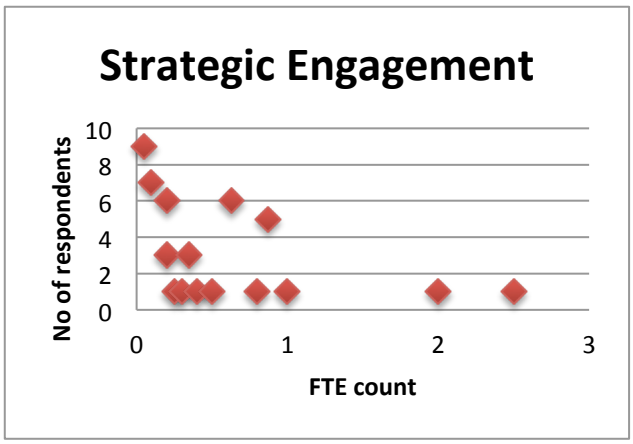
A further breakdown of FTE count against specific activities follows, however the reader should note that some approximations have been applied – e.g. “0.50 to 0.75 FTE” on the original questionnaire has been transformed to 0.63 FTE.

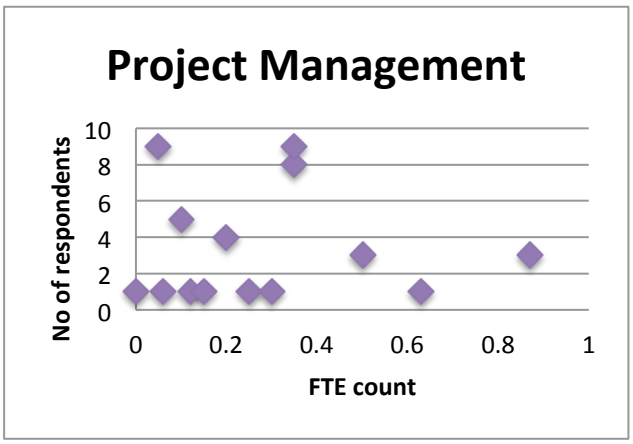


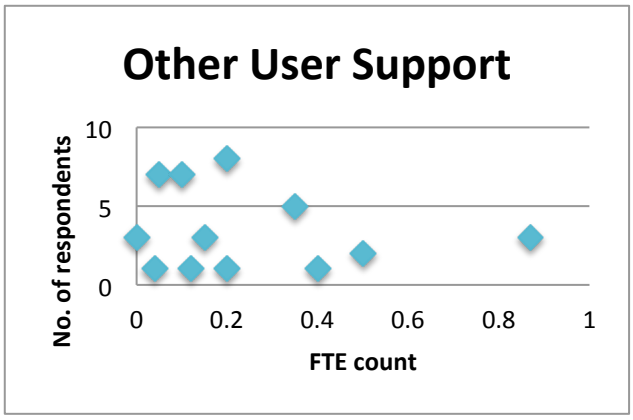



Application Support
No of respondents
8
6
4
2
0
0
1
2
3
4
5
FTE count


User Training
No of respondents
10
8
6
4
2
0
0
0.5
1
1.5
FTE count


Strategic Engagement
No of respondents
10
8
6
4
2
0
0
1
2
3
FTE count


Project Management
No of respondents
10
8
6
4
2
0
0
0.2
0.4
0.6
0.8
1
FTE count


Other User Support
No. of respondents
10
5
0
0
0.2
0.4
0.6
0.8
1
FTE count

11. Do you offer your own training courses?

72% of respondents overall offered training courses, with several respondents also noting that they promoted national level (e.g. ARCHER) training courses to their service users.

Do you offer your own training courses?		
Yes	36	72%
No	10	20%
Under development	4	8%

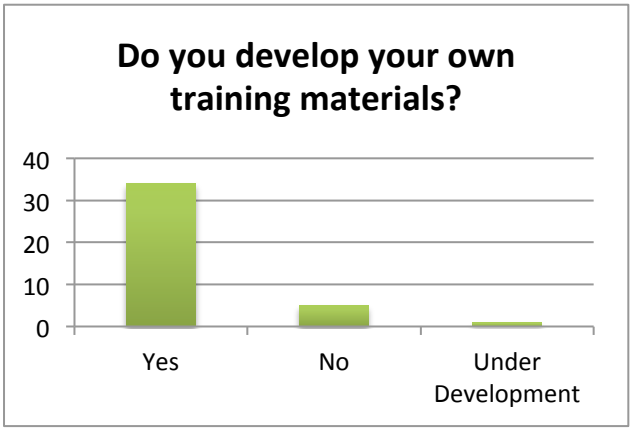
12. URL of training courses website if applicable:

http://www2.le.ac.uk/offices/staff-development/courses/it/hpc
https://www.wiki.ed.ac.uk/display/ecdfwiki/Courses+and+Events
http://www.shef.ac.uk/wrgrid/training
http://wiki.rac.manchester.ac.uk/community/Courses
https://apollo.hpc.sussex.ac.uk/HPCWiki/HPC
http://www.stfc.ac.uk/hartree
http://software-carpentry.org/bootcamps/index.html
https://apollo.hpc.sussex.ac.uk/HPCWiki/HPC
http://www.archer.ac.uk
https://hec.wiki.leeds.ac.uk/bin/view/Documentation/TrainingInformation
http://www.cardiff.ac.uk/arcca/services/events/index.html
http://www.bris.ac.uk/acrc
http://www.hpcwales.co.uk/skills-and-training

13. Training Materials – do you produce your own?

85% of all respondents indicated that they produced their own training materials, although several respondents noted that they were exploring ways of sharing training documentation or customizing existing course materials.

Do you develop your own training materials?		
Yes	34	85%
No	5	13%
Under Development	1	3%



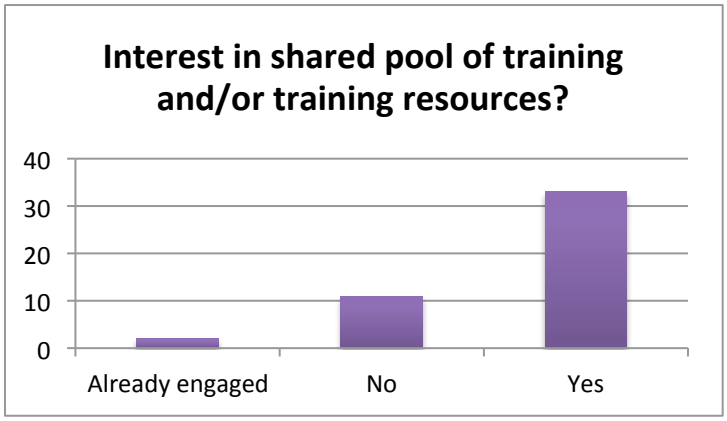
14. URL of training materials website if applicable:

https://www.wiki.ed.ac.uk/display/ecdfwiki/Courses+and+Events
http://wiki.rac.manchester.ac.uk/community/Course
https://apollo.hpc.sussex.ac.uk/HPCWiki/HPC
http://www.egi.eu/services/training_marketplace/
http://www.stfc.ac.uk/hartree
https://www.hpcavf.uclan.ac.uk/wiki/index.php/User_Guide
https://apollo.hpc.sussex.ac.uk/HPCWiki/HPC
http://www.archer.ac.uk
https://hec.wiki.leeds.ac.uk/bin/view/Documentation/TrainingInformation
http://www.cardiff.ac.uk/arcca/services/events/index.html
http://www.egi.eu/services/training_marketplace/
http://www.bris.ac.uk/acrc
http://www.hpcwales.co.uk/skills-and-training
http://www.ucl.ac.uk/isd/staff/research_services/research-computing/services/training

15. Are you interested in talking to SSI about contributing to a shared pool of training and/or training resources?

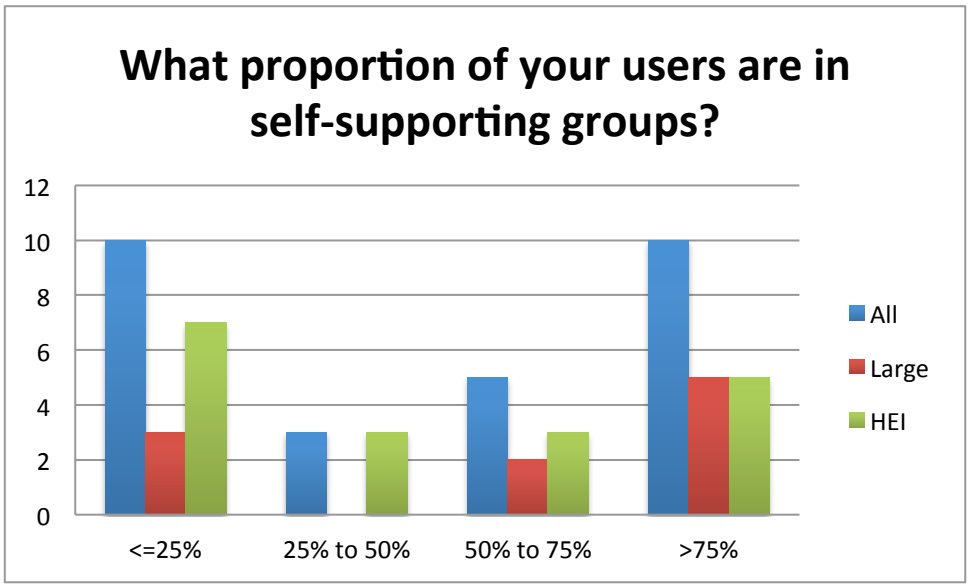
72% of all respondents were interested in engaging with the Software Sustainability Institute to develop a shared pool of training and/or training resources. 4% were already engaged with this process.

Interest in shared pool of training resources?		
Already engaged	2	4%
No	11	24%
Yes	33	72%



16. What proportion of your researchers are in self-supporting research groups?

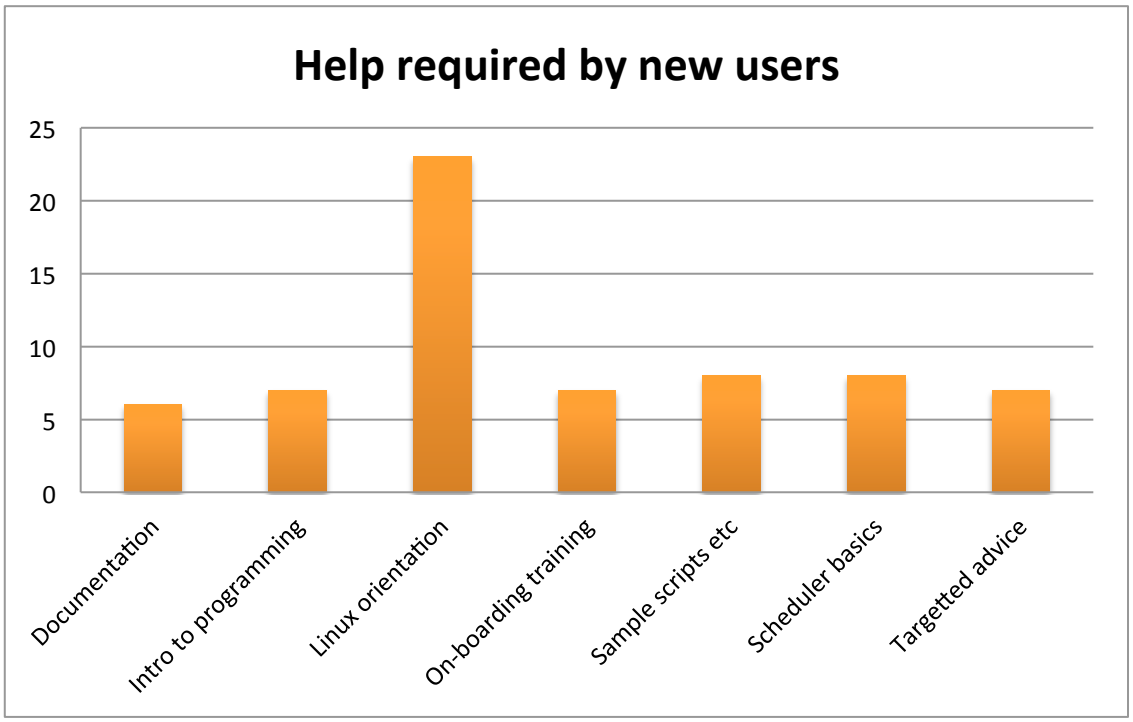
Both large and specialist services and HEIs reported that a high proportion of their users were working in self-supporting research groups, however 36% of respondents overall had a significant number of users that were not part of a larger group.



17. What sort of help does a new person typically need to use your systems?

Linux orientation was by far the most common support requirement for new users, with 35% of all respondents noting this as a key area. 11% of respondents stated that new users often required an introduction to programming before they could make use of the service.

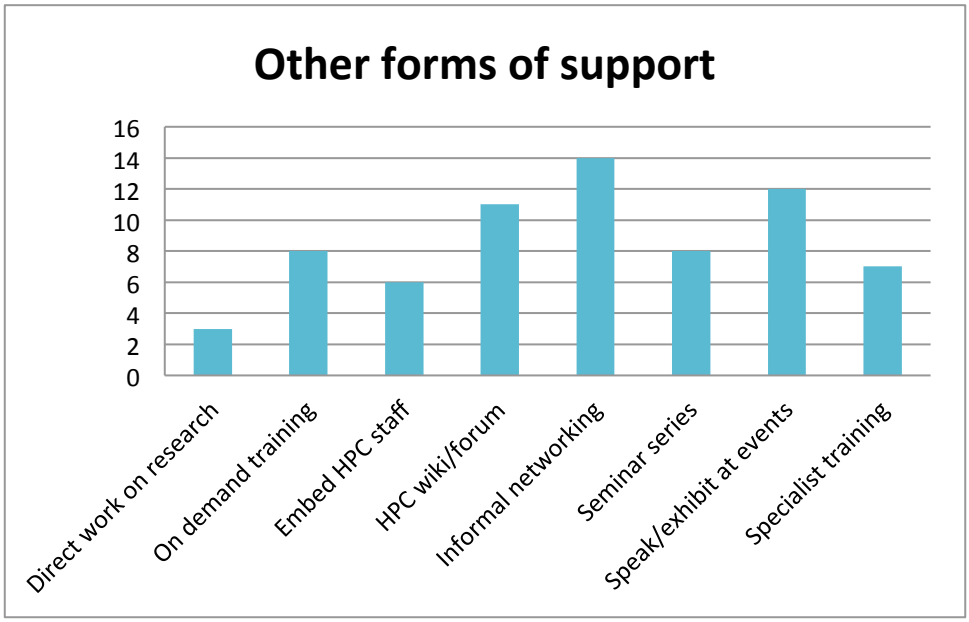
Help for new users		
Documentation	6	9%
Intro to programming	7	11%
Linux orientation	23	35%
On-boarding training	7	11%
Sample scripts etc	8	12%
Scheduler basics	8	12%
Targeted advice	7	11%



18. Other forms of support

Respondents were asked about other forms of user support offered as part of their HPC service provision. Several noted that they promoted courses available through the likes of EPCC, NAG, the European Grid Initiative and WLCG (Worldwide LHC Computing Grid), or through Doctoral Training Centre provision.

Other forms of support		
Direct work on research projects	3	4%
On demand training	8	12%
Embed HPC staff	6	9%
HPC wiki/forum	11	16%
Informal networking	14	20%
Seminar series	8	12%
Speak/exhibit at events	12	17%
Specialist training	7	10%



19. Are you able to provide HPC case studies for (e.g.) the HPC-SIG website, RCUK or InnovateUK (TSB)?

28 respondents indicated that they were able to do this.

20. Do you have a Research Data Management policy?

27 respondents had an RDM policy.

21. URL of Research Data Management Policy if applicable:

http://www.lboro.ac.uk/research/offcampus/ResearchDataManagementPolicy-Draft.pdf
http://www2.le.ac.uk/services/research-data
http://www.ed.ac.uk/schools-departments/information-services/about/policies-and-regulations/research-data-policy
http://www.calendar.soton.ac.uk/sectionIV/research-data-management.html
https://www.dur.ac.uk/research.office/research-outputs/
http://www.shef.ac.uk/ris/other/gov-ethics/grippolicy/practices/all/rdmpolicy
http://www.library.manchester.ac.uk/ourservices/research-services/rdm/policy/
http://www.sussex.ac.uk/library/research/researchdatamanagement/
https://www.stfc.ac.uk/1386.aspx
http://www.nerc.ac.uk/research/sites/data/dmp.asp
http://www.sussex.ac.uk/library/research/researchdatamanagement/
http://library.leeds.ac.uk/research-data-management-policy
http://n8hpc.org.uk/research/gettingstarted/filemanagement/RDM/
https://www.liv.ac.uk/intranet/media/intranet/researchdatamanagement/research-data-management-policy.pdf
http://www.cardiff.ac.uk/insrv/researchdata/managingdata/index.html
http://www.bris.ac.uk/acrc
http://www.esrc.ac.uk/about-esrc/information/data-policy.aspx
http://www.kcl.ac.uk/college/policyzone/assets/files/research/Research_Data_Management_Policy_v12_June_2013.pdf
http://gap.lancs.ac.uk/policy-info-guide/5-policies-procedures/Documents/SEC-2013-2-0776-Research-Data-Policy.doc
http://www.arcs.qmul.ac.uk/docs/policyzone/118815.pdf
http://www.ucl.ac.uk/isd/staff/research_services/research-data/researchdata/uclresearchdatapolicy
http://www.bath.ac.uk/research/data/

22. What technologies have you deployed that can be regarded as being for Data Exploration ("Big Data") activities?

- Several respondents were running production iRODS, Hadoop and NoSQL databases. It was noted that EUDAT had provided guidance on iRODS deployment, although this needed updating to cover the most recent iRODS version 4 release.

Individual responses follow:

- HPC Research Storage (PB scale NAS);
- We are running two large data stores: 1.1PB and 2.5 PB respectively. We also offer the Virgo Database, which is partially a mirror of another site, but is being extended to hold much more data. We hold archives of data on tape (> .5 PB);
- GPFS storage, large memory systems (1-2TB RAM);
- A peta-scale Hierarchical storage system for research data (disk/tape);

- We offer large memory Data Analysis Workstations and a service to support (including visualisation). We also use our 'HPC' facilities for large scale data analysis. We have just installed an SGI UV2000 for a research group on campus (Farr Institute/Iain Buchan). Data analysis of the output of the next generation sequencing facilities in Faculty of Life Sciences;
- Many of the technologies we service are deployed for distributed data exploration. Storage element technologies such as Disk Pool Manager and dCache. Catalogues such as the gLite local file catalogue. The File Transfer Service (FTS). Our user communities make use of data exploration tools such as ROOT and GEANT;
- IBM InfoSphere BigInsights Enterprise Edition, IBM InfoSphere Streams Developer Edition, Infosphere Data Explorer and Query Routing for Data Explorer, SPSS - Collaboration and Deployment Services and Deployment Manager, SPSS – Analytic Server Hadoop Pushback, SPSS – Modeller for data mining and Modeller Server, SPSS – Statistics Premium and Server, Cognos – Business Intelligence Architect and User, IBM Content Analytics production;
- See the separate Hartree Centre Response (IBM software Suite SPSS, Infosphere (Brightinsights, Streams, Data Explorer), Cognos, Content Analytics;
- None so far that are typical "Big Data" as defined by a Google (Hadoop) Map-reduce style, embarrassingly parallel, computation;
- Typically we deal with large volumes of data in an "unstructured" (to use the wider IT industry term to mean non RDBMS) manner. Of course, the files are typically model output in HDF5 or NetCDF which are rich in metadata;
- We have over a petabyte of data provided to the HPC and general computing environment via a parallel filesystem (GPFS);
- Looking at Hadoop on Cray XC30, SPRINT parallel R installed, genomic analysis using HPC, EUDAT stack on RDF, e.g. iRODS;
- UK-RDF + associated data analytics cluster (native setup, virtual data appliance facility, fast wide links to RDF disks) + data mover nodes (connection to PRACE 10 Gb/s pan-european network, lightpath to JASMIN), experimental Amdahl-balanced data exploration machine (EDIM1) based on Intel Atom processors and solid-state disks, SGI UV2000 + Trusted Edge + Infinite Storage Gateway;
- We have various activities in this area from pure computer science research right through to student projects. We currently have a couple of small [Less than 200TB] Hadoop clusters. We fully expect the data mining/analytics and data driven science to increase over the next few years and are exploring how best to support and integrate it with our existing systems;
- JANET Lightpath to RAL;
- Shared memory machine for Farr institute;
- 2TB SMP system and two remote visualisation nodes;
- Some HPC activities are of their nature "big data" and we accommodate these as and when needed. 2 examples are ½ Pbyte of storage for Space and Atmospheric Physics and SGI UV large shared memory and .75 Pbyte attached storage for genomics;
- We are currently exploring funding opportunities for a remote visualization capability, which would allow large data sets to be manipulated and visualized on the cluster directly. This is something our industrial users have expressed interest in due to connectivity issues;
- Shared memory, also MIC acceleration (Xeon Phi).

23. Other Information

Respondents were encouraged to provide any other information that might assist the e-Infrastructure Leadership Council and related bodies in their deliberations. Individual responses follow.

- GRIDPP is primarily an infrastructure provider;

- The HPC systems and services run by STFC support STFC staff, national, international and industrial activities;

- ARCCA are intimately involved in supporting both the HPC Wales and GW4 initiatives in HPC. The ARCCA datacentre is home to Raven and 10,000 cores of HPC Wales, while hosting equipment for other Departments in the University. HPC Wales vision is to provide state-of-the-art HPC capability, technology, infrastructure and facilities on a pan-Wales, pan-sector basis, to deliver research innovation, high-level skills development and transformational ICT for wider economic benefit;

- At HPC Midlands we have made significant inroads in developing standardised terms and conditions, a Service Level Agreement, and an intellectual property and information assurance regime that meets the requirements of firms ranging from micro-SMEs to major corporations. We believe there is scope for taking this work forward as a community led initiative to develop a national framework for e-Infrastructure services. Our experiences working with both academia and industry suggests that this framework will significantly reduce the friction of engaging with e-Infrastructure facilities. We are actively pursuing this opportunity with JISC and other e-Infrastructure centres of excellence.

Appendix G – Hardware: The Survey Questions

UK National E-Infrastructure Inventory 2014 - System Details

Q1. Organization name?
Q2. Organizational unit?
Q3. Respondent's email address?
Q4. Respondent's job title?
Q5. System name?
Q6. External IP address or FQDN?
Q7. URL for the website of the system or overall service?
Q8. Top 3 research areas the system is used for?

Hardware Specifications

Q9. Total number of processor cores in the system?
Q10. Number of compute nodes?
Q11. Number of processor cores per compute node?
Q13. RAM per core (Gigabytes)?
Q14. Compute node processor specification?
Q15. How many GPU equipped nodes does the system have?
Q16. How many Xeon Phi equipped nodes does the system have?
Q17. How many "fat" nodes does the system have, i.e. >=8GB RAM per core?
Q18. Does the system have a dedicated Visualization capability?
Q19. Interconnect Switch Fabric?
Q20. When was the system commissioned?
Q21. When will maintenance for the system terminate?

Storage

Q22. Describe the storage component of the system.
Q23. Total usable storage for HPC users?
Q24. What file system(s) do you use for shared storage?
Q25. Do you split system storage in terms of TB between fast, tertiary, archive storage?
Q26. Number of registered users?

Performance and Connectivity

Q27. Theoretical Peak Performance (Tflop/s)?
Q28. Node to Node Data Rate (Gbit/s)?
Q29. Average node to node latency (Microseconds)?
Q30. Typical CPU load as a % of overall system?
Q31. Peak CPU load due to industrial use over the last twelve months?
Q32. Peak inbound sustained data transfer rates?
Q33. Peak outbound sustained data transfer rates?
Q34. Connection Speed to JANET (Gbit/s)?
Q35. Is the bandwidth above dedicated for HPC service use?
Q36. Special connectivity requirements?

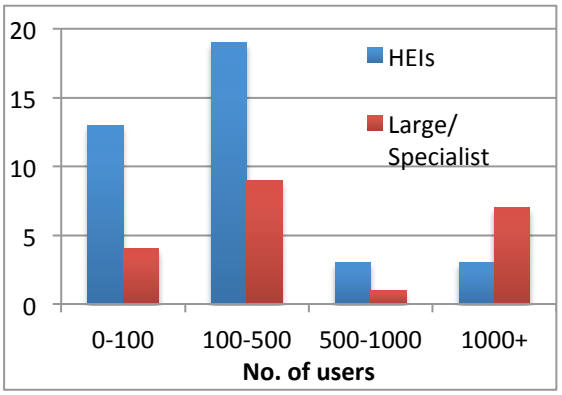
Operating Environment

Q37. What is the primary Operating System you use on compute nodes?
Q38. What is the primary Operating System you use on head/login nodes?
Q39. What scheduler(s) do you use?
Q40. Do you provide a Web Portal to your users?
Q41. Do you back up HPC user data?
Q42. Do you do have scheduled maintenance and if so how often?
Q43. Further Information

Appendix H – Hardware: Summary of the Survey Data

- We received 17 responses from Large and Specialist Projects;
- We received 35 responses from the 38 HEIs that are members of the HPC-SIG – See Appendix C for a full list of respondents;
- The summary of the responses is given below in the Tables and in written responses, which are to be found at the end of this section;
- The relative sizes of HEI local resources and the Large/Specialist systems (which includes GRIDPP and DiRAC resources at HEIs) in terms of resources and user base, is illustrated in the table and figure below.

	HEIs	Large/ specialist
Total no. of processor cores	91,122	458,068
Total theoretical peak performance (Tflops /s)	811	6501
Total amount of storage available to users (TB)	6,497	89,537
Total no. of users	13,225	15,282

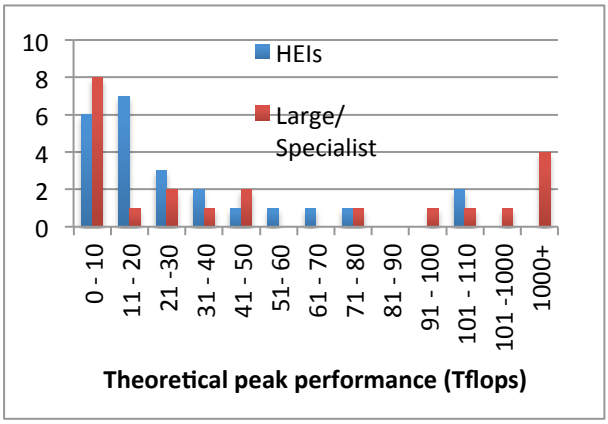


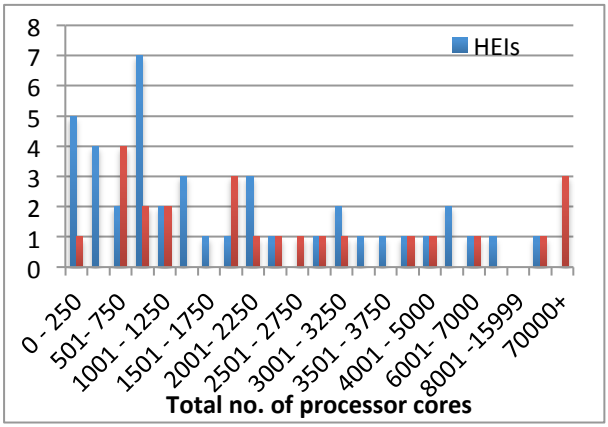
- The growing disparity reflects the investment in Big Data and HPC, which has largely gone to Large and Specialist Projects;
- An array of Big Data projects appear on the 2014 Survey. These are "people data projects" which include the Farr Institutes, the Admin Data Research Centres, the Admin Data Service, EMEDLAB, and the "physical science and engineering projects" such as the upgrade to JASMIN (NERC), the National Service Research Data Facility and the Hartree Centre's Energy Efficient Data Analytics Project;
- GRIDPP (UK node of the world LHC Compute Grid) is still the main Big Data and Compute Grid. In 2013 two new Cloud and Big Data Projects, the EMBL-EBI (Life Sciences, Informatics) and JASMIN (Earth Observation) were established and are fully functioning;
- Along with the EMBL-EBI and JASMIN (NERC), the "people data projects" projects above will be using virtualisation, e.g. secure Virtual Desktop Environments, to allow researchers to access and analyse data in a secure manner, as well as avoid large data flows between the Facility and the researchers laptop/PC – the users' CPU is being "brought to (or located at) the data";
- These projects bring sharply into focus requirements for AAAI and Information Security that need to be addressed by the Ne-I community via the RCUK Ne-I Group's Security and Access Working Group (Chair: Andrew Cormack, JANET);
- The works needed for improved AAAI and the use of virtualisation in many Large Projects are providing important building blocks for the emergence of real private clouds to support our research domains and improve our access issues;
- The Majority of the above projects come into service during 2014. All these services depend on good connectivity;

- New HPC services such as the new National Service, ARCHER, and large systems at Southampton and Bristol have come online in the last year;
- Very Large Energy Efficient Systems at the Hartree Centre have come on line and the NERC JASMIN Facility has been upgraded.

Compute capability

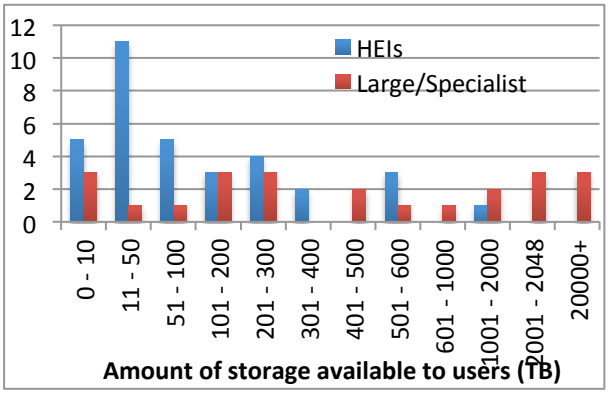
- The UK now has 3 PetaFlop systems available to UK academics, either by peer review or via paid access. These are ARCHER (National Service), Blue Joule (academics will need to be working with an industrial PI/Partner) and the DiRAC Blue Gene;
- There are now 10 systems available between 100 and 500 Tflops;
- The majority of systems are under 50Tflops and are based at HEIs;
- A number of HEIs have compute provision which is (albeit when crudely measured in terms of core count) on a par with large and specialist facilities;
- 97% of systems were running a Linux variant.





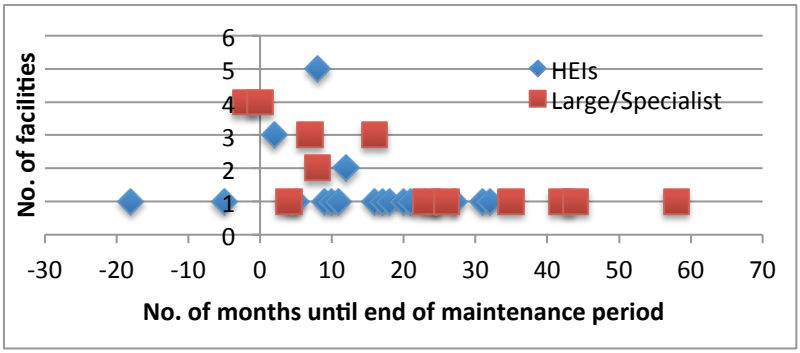
Storage

- There has been an increase in the number of services with > 1PB of storage. There are now 15 such services in the Large and Specialist systems. However only 2 HEIs have in excess of 1PB of storage;
- Highly Performing and Large, Storage and Clusters services are run in the main by National Centres and National Projects.



Sustainability

- A large proportion of systems would be due for renewal in the next 12 – 24 months, and several institutions were



running systems off-maintenance.

Table: The hardware systems are given in the Table below which is split to into 4 sections, each representing a layer in the Branscomb Pyramid. Section 1 lists the Large and Specialist Systems. Section 2 gives the Regional Systems. Section 3 lists the UK Nodes Tier 1 and 2 nodes of GRIDPP, the UK part of WLCG. Section 4 lists the foundational layer, the HEI sector.

Organization name	Top 3 research areas the system is used for	Total number of processor cores in the system	Total usable storage for HPC users (TB)	Number of registered users	Theoretical Peak Performance (Tflop/s)
1. Large and Specialist Services					
National Service - ARCHER	EPSRC, NERC	72192	19000	976	1645.7
National Service - Indy	CFD explosions, MD Chemistry	1536	46	110	14.3
National Service - Hydra	Numerical Algorithms	284	3	50	
DIRAC@Cambridge - COSMOS	Cosmology, astrophysics, space science	1856	200	200	76.0
DiRAC@Cambridge - COSMOS	Cosmology, astrophysics, space science	732	60	200	9.0
DIRAC@Cambridge - Darwin	Cosmology, astrophysics, nuclear physics, particle physics	11328		100	225.6
DIRAC@Durham - Cosma4	Cosmology, astrophysics	2976 intel cores, 4096 GPU cores	1100	188	31.8
DiRAC@Durham - Cosma5	Cosmology	6720	2400	227	139.8
DiRAC@Edinburgh - Bluegene/Q	Particle physics & astronomy	98304	1000	171	1260.0
DiRAC@Leicester - complexity	Cosmology, astrophysics, nuclear physics	4352	750	0-100	90.5
STFC Hartree – Blue joule	CFD, Materials Chemistry, Weather & Climate Modelling	131072	25000		1800.0
STFC Hartree – Blue Wonder	CFD, Materials Chemistry, Weather & Climate Modelling	18400			500.0

STFC Hartree – Novel Cooling	CFD, Materials Chemistry, Weather & Climate Modelling	1920	216		N/A
STFC Hartree	CFD, Materials Chemistry, Weather & Climate Modelling	FPGA system	112		N/A
STFC Hartree	CFD, Materials Chemistry, Weather & Climate Modelling	Nextscale + ARM	100		N/A
STFC Hartree - BigData	CFD, Materials Chemistry, Weather & Climate Modelling	1056	1000		N/A
STFC SCARF	Computational Chemistry Plasma Physics, Processing Satellite images	3500	120	200	42.0
STFC GPU for MRC	3D tomography reconstruction	192	60		N/A
STFC Columbus	Computational Chemistry	512	50	400	5.0
STFC Columbus2	Computational Chemistry	512	50	400	10.0
STFC JASMIN/CEMS	Climate Change, weather modelling, Satellite Data analysis	648	5500	14000	6.0
STFC JASMIN2	Climate Change, weather modelling, Satellite Data analysis	3,648	6900		73.0
STFC GRIDPP Tier 1	Particle Physics	7000	20000	1200	See below
STFC Visualisation Space	Space applications				N/A
STFC Visualisation	Analysis of ISIS, Diamond and Laser data	32	10		N/A
The Francis Crick Institute (incl EMEDLAB)	Next Generation sequencing. High resolution high through put biomedical imaging. Macromolecular structure	240	500	0 - 100	11.0
The Genome Analysis Centre	Genomics & Computational Bioscience	768	20	100 - 500	1.7
The Genome Analysis Centre	Genomics & Computational Bioscience	768	2000		2.0
The Genome Analysis Centre	Genomics & Computational Bioscience	2560	2000		6.6
The Genome Analysis Centre	Genomics and Computational Bioscience	3000	2000		6.3
WTCHG, Oxford	Genomics	1152		80	9
WTCHG, Oxford	Genomics	248		10	1
WTCHG, Oxford	Genomics	912		130	14
WTCHG, Oxford	Genomics		1200		N/A

Sanger Institute	Genomics	6048	3000	200	47.1
Sanger Institute	Genomics	2368	1000		17.7
Sanger Institute	Genomics	2232	1000		15.9
Sanger Institute	Genomics	1016	2000		6.9
Sanger Institute	Genomics	3200	5000		15.9
Sanger Institute	Genomics	2400	5000		6.9
National Oceanography Centre	Marine systems modelling Physical Oceanography Marine Geology and Geophysics	1152	225	0 - 100	21.6
2. Regional Systems					
HPC Wales Cardiff	Advanced Materials & Manufacturing, Life Sciences, Creative Industries and Energy & Environment	2072	171	1521	21.8
HPC Wales Cardiff	ditto	6400	171		133.1
HPC Wales Swansea	ditto	4096	338		95.0
HPC Wales Swansea	ditto	2304	338		47.9
HPC Wales Aberystwyth	ditto	648	10		6.9
HPC Wales Bangor	ditto	648	10		6.9
HPC Wales South Wales	ditto	648	10		6.9
HPC Midlands Loughborough	Aeronautical and Automotive Engineering Mechanical and Manufacturing Engineering Materials	3008	120	0 - 100	48.0
N8 Research partnership	Materials/chemistry, molecular modelling, Atmospheric chemistry.	5536	174	325	115.0
N8 Research partnership	Health Informatics	256	60	5	5.0
STFC/SES GPU	Life Sciences	1008	135	100	
SES Southampton	Life Sciences, Engineering, Physics	12000	250	500-1000	106.0
Strathclyde /Archie-West	CFD (Mechanical & Aeronautical Engineering) Molecular Dynamics Plasma Physics	3920	150	100 - 500	36.3
Warwick/Midplus		6036	200	100 - 500	65.0
The Institute of Cancer Research	Next generation sequencing, image	2000	2700	0 - 100	40.0

	processing, molecular dynamics.				
3. GRIDPP Services	**LHC Physics analysis (for the ATLAS, CMS and LHCb experiments)**	**43975**	**24234**	**8000+ via Grid certificates and membership of Virtual Organisation. Authorization is controlled by VOMS.**	**GRIDPP total HEPSPEC06 score is 377,000 which equates to ~43,100 Intel 2650 HT cores, which equates to 345Tflops.**
RAL Tier 1		11324	13049		90.6
UKI-LT2-Brunel		1492	592		11.9
UKI-LT2-IC-HEP		3216	2026		25.7
UKI-LT2-QMUL		3536	1680		28.3
UKI-LT2-RHUL		1744	728		14.0
UKI-LT2-UCL-HEP		240	159		1.9
Lancaster		2625	1107		21.0
Liverpool		952	544		7.6
Manchester		2630	1032		21.0
Sheffield		848	360		6.8
DURHAM			42		0.0
ECDF		5792	355		46.3
GLASGOW		4136	1405		33.1
EFDA JET		192	1.5		1.5
Birmingham		816	315		6.5
Bristol		580	117		4.6
Cambridge		220	278		1.8
Oxford		1384	708		11.1
RALPP		2056	1530		16.4
Sussex		192	70		1.5
4. HEI Services					
Cardiff	EPSRC, BBSRC and NERC	2048	275	400	42.6
Cardiff	EPSRC and BBSRC	864	275	400	9.7
Durham	Condensed Matter Molecular Dynamics Computational Fluid Dynamics	3600	35	100 - 500	35.0
Imperial College London	Engineering, CFD, CMP	16000	1000	1000	192
Imperial college London	Bioinformatics	5272	600	100 - 500	63.3
Imperial college London	Genomics Computational Chemistry	384	500	0 - 100	4.6
King's College London	Physics, Chemistry,	1464	87	55	58.0

	Informatics				
Lancaster	High Energy Physics / Molecular Dynamics / Maths & Stats	2200	31	200	20.0
Liverpool	Computational Chemistry, Human Anatomy (gait analysis), Engineering (CFD)	1200	48	127	8.0
Liverpool	Aerospace CFD, Computational Chemistry, Biochemistry	2448	100	39	40.0
Loughborough	Aeronautical Engineering Materials Civil Engineering	1972	50	100 - 500	18.0
QMUL	Engineering & Material Science, Astrophysics, Biological & Chemical Sciences	3188	220	692	38.3
Queens Belfast	Speech and Image Vision Systems Civil Engineering Biology	256	29	0 - 100	4.5
Queens Belfast	Computational Chemistry Plasma Physics Cancer Research	928	51	100 - 500	11.1
Aberdeen	Engineering - CFD Life Sciences - Bioinformatics	608	56	100 - 500	10.0
Bath	Chemistry, Physics, Maths	824	23	50	9.0
Birmingham	Engineering Bioinformatics Computational Chemistry	1000	150	100 - 500	18.0
Bristol	Soc&Com, Med, Climate, ElecEng	10000	740	800	240.0
UCLAN	Astrophysics Computation Engineering Computational Physics	544	45	0 - 100	5.8
UCL Computer Science	CS, Medical Imaging, Bioinformatics	3500	2000	400	42.0
UCL	Biosciences, Physics, Chemistry	7808	380	580	108.0
East Anglia	Climate Research, Machine Learning, Molecular Modelling	4148	170	500	65.0
Edinburgh	Physics, Informatics, Geosciences	3144	281	1000	29.0
Exeter	Astrophysics, Hydrology, Applied Mathematics	2184	69	98	24.5
Glasgow	Semiconductor	1360	34	0 - 100	16.3

	Device Modelling				
Glasgow	Semiconductor Device Modelling Computational Fluid Dynamics Optoelectronics	1256	22	0 - 100	15.1
Glasgow	Computational Fluid Dynamics	188	4	0 - 100	N/A
Glasgow	Electronic System Design	320	10	0 - 100	N/A
Huddersfield	Engineering - CFD, FEA;	160	24	200	0.8
Huddersfield	Applied Sciences - MD, Chemistry, Physics	260	24	200	2.5
Huddersfield	ditto	8	1	4	N/A
Huddersfield	ditto	216			12.2
Huddersfield Hadoop	ditto	40	2	200	N/A
Huddersfield Condor	ditto	3500	24	200	
Leeds	Engineering, Maths & Physical Sciences, Environment	1952	110	567	22.0
Leeds	Astrophysics, Atmospheric Chemistry, Materials	3360	110	567	35.0
Leeds	As ARC1, Biomedical sciences	3040	170	209	63.0
Leicester	Physics (Astrophysics and Earth Observation Science) Engineering Economics	3818	344	1000	75.0
Manchester	CFD Chem Eng / Molecular Dynamics Bioinformatics	5488	151	500 -1000	101.3
Nottingham	Physics & Astronomy, Chemistry, Engineering	2656	550	210	47.0
Sheffield	Mechanical Engineering, Physics, Chemistry	912	45	100 - 500	15.0
Sussex	Physics (Astronomy,Cosmology,Particle), Engineering (CFD), Informatics (Computational Neuroscience, Adaptive systems, Natural Language)	2852	594	120	35.0
Southampton	Life Sciences, Engineering, Physics	12200	850	800	260.0
Oxford	Structural Biology, Biochemistry / Chemistry, Materials	640	600	1000	6.0

Oxford	Ditto	640	600		4.0
Oxford	Ditto	1344	600		30.0
Oxford SMP	Ditto	64	600		N/A
Oxford GPU 1	Ditto		600		N/A
Oxford GPU 2	Ditto		600		N/A
Cranfield University		1280	34	100-500	19.8
Strathclyde	CFD, Molecular Dynamics, Plasma Physics	1000	135	0-100	12.0
Strathclyde	Ditto	64	135		0.5
Strathclyde	Ditto	3408	159		36.3
St Andrews	MHD, astronomy, chemistry.	2424	150	100-500	28.0

Written Responses

The EMBL European Bioinformatics Institute

System Name: EMBL-EBI
HEI: EMBL-EBI
Contact: steven.newhouse@ebi.ac.uk
Total Cores: 40,000
Storage: 46PB (various systems - NAS/SAN/SSD/Cluster/Tape)
JANET: 10Gb multiple links
Areas: Life Science

The Farr Institutes

Farr @ Cipher

Farr investment – the additional activities in addition to buildings that have resulted from the Farr investment that are centre specific. We have spent £2.5m on Farr associated technology, including the procurement of Research Data Appliances (RDAs) for Wales and SW England and UKSeRP, facilities to be operational by June 2014.

Farr @ Scotland

After successful negotiations, £1.3m has been spent on IT servers and storage to establish a high powered Farr computing infrastructure and analysis environment. This is being hosted by Edinburgh Parallel Computing Facility, one of Europe's leading supercomputer facilities. Information governance and security specialists are working with IT specialists to ensure that the design and services run on the infrastructure meet information governance and security requirements as a pre-requisite to consider hosting new national datasets in the environment.

Farr @ HeRC

HeRC has invested in computational infrastructure and secure rooms for a data safe haven at Manchester, including a direct NHS N3 connection (£0.46m) and additional HPC capacity in collaboration with the N8 HPC Service at Leeds (£0.28m). To support the CoOP theme we have created a pool of smartphones, tablets and activity/sleep wearable trackers to enable preliminary data collection, feeding into new grant applications (£0.17m). Through the Northern AHSNs we have invested in infrastructure projects for integrating primary and secondary care data (Liverpool, Morecambe Bay, Newcastle), and awarded a tender to NWeH for the Farsite tool for clinical trial feasibility and recruitment with these new data sources for the AHSNs (£0.7m). We have established infrastructure for mortality assessment across West Yorkshire; funded the infrastructure for a regional PROMs collection service and enabled data from TPP ResearchOne for clinical trial protocol feasibility and recruitment (£0.4m across three projects).

Farr @ London

We have focused our investment where it can best facilitate the flow of data from clinical to research environments for processing, analysis, storage and publication. Where possible we are bridging infrastructure across the Farr London partners. For example we have implemented a virtual machine infrastructure to make pseudonymised clinical data collected from over 36,000 cardiovascular patients at Barts accessible to Farr researchers (100K), which builds on a secure data transfer layer enabling data to flow between UCLP NHS Trusts and Farr (174k). The processing requirements for analysing complex EHR datasets are increasing exponentially with the introduction of genomic and image data and the linkage of multiple sources. To meet this growing demand we have invested in significantly enhancing our compute and data storage capacity. We have purchased >2PB of storage space (£550K) in order to expand our current infrastructure and make it more resilient and have added a further 1600 cores to UCL's existing high performance computing cluster. A secure data centre has been installed to host the purchased equipment (442K) with additional cross-campus network enhancements (£107k). Finally we have established a clinical information systems training platform by purchasing training licences for EMIS and CERNER systems in order to facilitate training and capacity building (£189K), with the aim of embedding health informatics training in the medical student curriculum, starting in autumn 2014.

Admin Data Service and Admin Data Research Centres

The Administrative Data Research Network (ADRN) consists of four Administrative Data Research Centres (ADRCs) - the ADRC for England led by the University of Southampton, the ADRC for Northern Ireland led by Queens University Belfast, the ADRC for Scotland led by the University of Edinburgh, and the ADRC for Wales led by Swansea University.

The Administrative Data Service (ADS) will be led by the University of Essex.

The ADRC-England reported to the Survey that they will be providing state of the art secure facilities to the research community and be delivering the first services in May

- Safe settings compliant with Impact Level 2 and 3 at Southampton and in partnership with the Farr and at Impact level 4 at the Jill Dando Institute. In total we aim to have over 30 seats a locations in London and Southampton and through a secure and scalable Virtual Desktop Environment, enable ~100 simultaneous researchers to remotely work on research projects categorised as Impact level 2.

Appendix I – Use of software for scientific computing in the UK

Neil Chue Hong, Software Sustainability Institute

Executive Summary

Scientific software remains a critical part of the UK's e-Infrastructure. Compared to the results of the 2013 survey, the responses from the 2014 survey show that the main issues and concerns remain. As recognition and provision become more widespread, there is convergence on the importance of *basic software engineering training for researchers, research software engineer career paths, and sustaining a critical mass of software expertise*.

Scope

The UK Scientific Computing Software survey went out to major scientific consortia with an interest in scientific computing (e.g. the CCPs, and groups such as GRIDPP, NCAS, Virgo) and to research institutions with scientific computing infrastructure (through the HPC-SIG). A total of 24 responses were received split between 6 scientific consortium responses (including 3 CCPs), and 16 institutional responses (there were multiple responses from some consortia and institutions).

As expected the number of people in the organisation reported by institutional respondents was generally >250, and the number of people in scientific consortia varied but demonstrated strong modes at 11-50 or >250 people.

Software Packages and Infrastructure

The software packages reported in use varied widely at institutions. However some packages were available on the majority of e-Infrastructure: Matlab, R, and Python/NumPy/SciPy along with the Intel and GNU compilers. Other common software (4+ responses) included:

Gaussian, VASP, and CASTEP. Each had a wide set of in-house / self-written software primarily coming from research teams using the facilities, including CCP flagship software.

Software packages provided by scientific consortium tended to reflect the specialist tasks that the consortium members undertook in the conduct of their research: commercial reconstruction (VGStudio) and quantification (Avizo) software; microscope software from FEI and Gatan; simulation frameworks (COPASI), information analysis platforms (Spotfire, knime, Pipeline Pilot, Cytoscape); computational frameworks (MATLAB, Mathematica, R) as well as open source compilers (GCC, Ruby, Java, Python), workflows systems (Galaxy, Taverna), parallelisation frameworks (OpenMPI, MPICH) and software development infrastructure and tools (Git, mercurial).

A new question in this year's survey asked about programming languages available on the UK e-Infrastructure. The most popularly available programming languages were C++, Fortran, and Python. C, Java, MatLab and R were also commonly available.

Looking at software infrastructure provision, as was reported in 2013, the majority (over 50% of respondents) provided knowledge repositories (e.g. wikis), website hosting and mailing lists and version control repositories. However again few (less than 1/4) provided project management tools or build and test / continuous integration environments, suggesting that infrastructural support for high-quality, long-term scientific software development is still lacking.

Training and Guidance

The large majority of respondents provided some form of training to people in their organisation, on topics which are part of the "general curriculum" for a researcher using scientific software.

The majority provided specialist training in particular research software tools and packages, and on specific programming languages. As expected (due to their involvement in HPC-SIG), 11 of the 17 institutional respondents provided training on parallel and distributed programming. About half the respondents (distributed evenly across categories) provided training in basic programming and software engineering skills, which shows that there are still gaps to fill here as arguably everyone should now have access to this sort of training.

There is still limited access to training on advanced software engineering skills, numerical algorithm development and the understanding of new developments in computer hardware. Availability of training on how to apply computational techniques and data curation and management has improved slightly since last year's survey but is still available at only ⅓ of the organisations surveyed.

There is widespread awareness and use of nationally provided software training resources such as the National HPC Service (HECToR) training courses (provided by NAG), EPCC training courses, PRACE Advanced Training Centre courses, and the Software Carpentry workshops coordinated by the SSI. HPC Short Course Centre courses were perhaps underrepresented as only 50% of respondents were aware of them.

When asked to highlight the areas considered to be under-provided in terms of training and guidance to support software use and development, there was a clear priority: basic software engineering skills (with Software Carpentry mentioned by name several times).

Scientific consortia felt that the following topics were underprovided:

- Basic software engineering skills;
- Using clusters / Parallel and distributed programming;
- Data analysis techniques;
- Choosing and using software appropriately;
- Bid writing, which includes software.

Whereas institutions highlighted:

- Basic software engineering skills;
- How to apply computational techniques as research tools;
- Data analysis, data curation and management;
- Courses tailored for X-informatics subjects;

- Parallel programming.

It is clear that there are still areas of under-provision of training. However it appears that more and more organisations are cognisant of the requirement for basic software engineering, and indeed are specifically providing or requesting Software Carpentry. The remaining issue is the under-resourcing and under-valuing of this training, which is often expected to be provided free at the point of delivery.

Software Challenges

When asked what they would like to see more recognition of, the clear standout was *career paths for developers of scientific software*. From scientific consortia, optimisation methods, recognition of software as a research object / credit for software, and recognition of software planning / sustainability in grants by investigators and reviewers were raised as issues. From institutions, reproducibility / correctness of results was highlighted, and sustainability of software, more robust Linux installation processes, general training for staff and postgrads, licensing, software as a research output and new technology expertise were also mentioned.

Finally, when asked about the single most pressing research software challenge they faced, it is clear that there are differences between the responses from the scientific consortia and the institutions.

Scientific Consortia highlighted:

1. Long term sustainability of software;
2. Issues related to management of rewriting code when targeting new hardware;
3. Time and staff to develop software and community.

Contrast with the 2013 responses:

1. Big data / streaming data;
2. Not enough software expertise;
3. Sustainability / long-term stability of funding for software used.

Institutions highlighted:

1. Recruiting and retaining people with appropriate scientific software development skills;
2. Embedding and dissemination of best practice in scientific software development between disciplines;
3. Better awareness of packages.

Contrast with the 2013 responses:

1. Cost of licenses;
2. Researchers lack software skills / Institutions lack resources to enable researchers to access or embed software expertise;
3. Development of codes / availability of libraries to take advantage of modern architectures;
4. Management and knowledge of a diverse range of software.

Therefore the underling headline challenge remains the same as last year: *ensuring a sustainable, critical mass of software expertise*.

Respondents

- Scientific Consortia
    - ASEArch CCP
    - CCP-EM
    - CCPi
    - DiRAC
    - ELIXIR-UK
    - MyGrid
    - SysMO-SEEK
- Research Institutions
    - Durham University (Institute for Computational Cosmology)
    - The Institute of Cancer Research, London
    - King's College London
    - National Oceanography Centre
    - Queens University of Belfast
    - STFC Centre for Environmental Data Archival
    - University of Bath
    - University of Bristol
    - University College London
    - University of Edinburgh
    - University of Glasgow (School of Engineering)
    - University of Manchester (academic researchers)
    - University of Manchester (IT Services)
    - University of Leeds
    - University of Liverpool (Advanced Research Computing)
    - University of St Andrews
    - University of York (IT Services)