

Cloud computing for research

final report

CC421D007-1.0

7 June 2010

Cover + intro + 60 pages

Dr Max Hammond
Dr Rob Hawtin
Dr Lee Gillam
Prof Charles Oppenheim

curtis+cartwright 

Curtis+Cartwright Consulting Ltd

Main Office: Surrey Technology Centre,
Surrey Research Park, Guildford
Surrey GU2 7YG

tel: +44 (0)1483 685020
fax: +44 (0)1483 685021
email: postmaster@curtiscartwright.co.uk
web: <http://www.curtiscartwright.co.uk>

Registered in England: number 3707458

Registered address:
Baker Tilly, The Clock House,
140 London Road, Guildford,
Surrey GU1 1UW

Executive summary

- 1 Curtis+Cartwright Consulting Ltd, working with the University of Surrey and Professor Charles Oppenheim, has been commissioned by JISC to investigate Cloud Computing for Research. This document is the final report, and is accompanied by a briefing paper which provides advice targeted at researchers.
- 2 The scope of this report is to consider Cloud Computing for research in the areas of compute and storage. Infrastructure as a Service (IaaS) and Platform as a Service are in scope, while Software as a Service is not.
- 3 This document (and the accompanying briefing paper) are intended to provide decision support, and not a single recommendation: critical thinking is still required from researchers and institutions as to what data storage or compute solution is most appropriate given functional requirements, budget, security, reliability, trust, *etc* as well as the cloud services currently on offer.

The current situation

- 4 There is currently an extremely rapid pace of change and innovation across the range of activities considered to be Cloud Computing. This covers not just the details of costs of services from specific providers, but also the range of services offered and the available tools for configuring and controlling cloud resources.
- 5 We present here a snapshot of the situation in April 2010. All readers are strongly urged to keep in mind the fast changing nature of the field: our descriptions of services will be out of date within months from publication, and entirely new ideas, technologies, business models, providers, strategies and risks will continue to appear.
- 6 There is a wide range of cloud vendors, although three make up the bulk of the market: Amazon Web Services (AWS), Google App Engine, and Microsoft Windows Azure.
- 7 There is a broadening range of academic research into the use of Cloud Computing for research, although as a proportion of research computing overall it is insignificant. At present, the vast majority of research computing using the cloud is undertaken on AWS, funded by Amazon.

Barriers to uptake

- 8 There is some resistance to change amongst academics, but this resistance is for the most part well considered and argued – it is not naïve obstructionism. There are significant barriers to the large-scale uptake of Cloud Computing for research, described within Section 3.
- 9 The core issue is that at present the costs of research computing are frequently hidden from the users of the service. Institutions typically provide some or all of the facility as an overhead, paid for from institutional or departmental budgets. These overheads include buildings, power, cooling, staff and administration. This creates an environment where very few researchers have a clear idea of the true costs of the resources they are consuming.
- 10 Cloud Computing provides an environment where many more of the costs are exposed to the users of the service, and so the service is more expensive *to the users*. Unless an

organisation has fully identified the costs of local provision, and understood how those costs would be changed by adoption of cloud systems, it is not possible to judge which approach is less expensive *overall*.

- 11 Cloud Computing may well provide some significant benefits for certain kinds of research (see below), but in the case of economic benefits, these accrue at an institutional or national level rather than to the researchers themselves. There may indeed be a disbenefit to the users in adopting cloud services. Without a fundamental shift in institutional (or Research Council) approaches to costing, this misalignment of benefits and costs is likely to preclude extensive natural adoption of cloud technologies for tasks which are currently well met by local provision.

Information assurance

- 12 There are a wide range of concerns about the information assurance aspects of cloud provision. Many of the issues that are raised are not well addressed within current local or institutional provision, but there is often a sense (rightly or wrongly) that local services are more trustworthy than external ones.
- 13 A common question is “who can access my data on the cloud?” This question has little meaning – “the cloud” is not an entity as such, and it is up to the user to decide which services to use, and each service has a distinct approach to data. Generally, services bought from cloud vendors are private services – your data (and other resources) belong to you and will not be accessible to others unless you make it so.
- 14 From a security (confidentiality) aspect, many potential users expect absolute guarantees that their data cannot be accessed without their authorisation, but it is never possible to give these guarantees. For example, it is reasonable to expect services to protect against common attacks, and to not release user data to the internet. But what about skilled and well-resourced attackers who might be targeting an organisation specifically? New vulnerabilities are constantly discovered in all elements of the internet, and until they are disclosed, they will be exploitable. What about legal requests from police or other organisations?
- 15 Regarding the availability and integrity of data (commonly referred to as “backup”, although the issues are broader), cloud provision provides just one more option in the range of services available to institutions or their researchers. Approaches must be considered on a case-by-case basis.

Contracts and liability

- 16 The contracts and Service Level Agreements (SLAs) currently on offer are typical agreements of their type. As such, they are one sided in favour of the service supplier and against the client. The style is very much lawyer-speak, there is a risk that HEIs, or non-specialist staff within them, will sign without realising all the implications. Sub-section 5.8 contains a detailed analysis of the contracts and SLAs on offer from the key vendors.
- 17 One particular issue to note is that these contracts are established under US, rather than any UK law, and require conflicts to be settled in US courts. Our analysis highlights some areas where the contracts may be illegal under UK law, but may be permissible under their own jurisdiction.

- 18 In summary, the services agree to supply a service to the clients for a fee, but only accept limited liability if things go badly wrong and they do not accept basic legal obligations on their operations.
- 19 Whether the contracts and SLAs offered are appropriate for any given research task will need to be considered on a case-by-case basis. It is particularly important that users consider the obligations that they have under the Data Protection Act and the Freedom of Information Act (see sub-section 3.6) Vendors are unlikely to negotiate changes for individual clients, but opportunities to enable more business with the UK HE sector as a whole may prove enticing.

Applicability to research computing

- 20 Cloud Computing services could theoretically fulfil any research computing requirement. However, two major areas have emerged as being currently **unsuitable** for migration to a cloud platform. However, this is an assessment made on current cloud provision only, and may well be subject to rapid change as vendors bring new services and offerings into production.
- **Large-scale data handling and processing** is currently likely to prove unaffordable due to charging models used by commercial cloud service vendors. This could potentially be negotiated, if the scale were large enough.
 - **Fine-grained/tightly-coupled parallel processing** for example using OpenMP or MPI is currently likely to suffer unacceptable performance overheads from virtualisation into a cloud.
- 21 That said, the management of research data is a significant current problem. Cloud provision provides opportunities for storage and management of research data, albeit with a new set of problems regarding sustainability, longevity, and cost.
- 22 We present a range of potential use cases (sub-section 4.4) where Cloud Computing could provide important new capabilities. We argue that it would be more beneficial to concentrate on these (and other novel cases), rather than displacing existing cluster and High Performance Computing (HPC) provision. Our list of potential use cases includes:
- Short timescale requirements;
 - Infrequent use and/or no desire to maintain infrastructure;
 - Dynamic scalability to larger capacity ('cloudbursting');
 - Transfer to commercial use;
 - Flexibility with system configuration and/or frozen system configuration;
 - Data hosting and backup;
 - Cloud-based research publications;
 - Ad hoc activities in support of research.
- 23 The JISC '*Using Cloud for Research: A Technical Review*' project presents a range of usage scenarios, which may provide approaches to meeting these use cases.¹
- 24 Although we made efforts to engage with research computing users within the Arts, Social Sciences and Humanities, the number of individuals active in this area is very limited, and the

¹ *Using Cloud for Research: A Technical Review*, (§3.2) Xiaoyu Chen, Gary B. Wills, Lester H. Gilbert, David Bacigalupo, May 2010.

majority of these are using workstation computing rather than large-scale provision. These users could potentially benefit from the flexibility of IaaS clouds to give them more of what they already have, without the significant challenges of migrating to Unix or Linux to operate on clusters or HPC machines.

- 25 It is clear that institutions should consider their approach to Cloud Computing – is it acceptable for individual researchers to use these services without institutional consideration of the risks? Many of the issues are similar to those around individual researchers buying and configuring their own local resources; who is responsible, and whose reputation is damaged if there is an incident?
- 26 Staff skills are unlikely to be significantly affected by any shift to Cloud Computing. The individuals who are involved in research computing at present are technically-minded, and cloud presents just one more technological opportunity. Cloud Computing is sometimes held up as a panacea which can allow non-technical users (and non-traditional research computing users from the humanities) to easily access powerful facilities. This is not the case, for IaaS and PaaS clouds at least, where significant technical expertise is required to provision and administer these systems.
- 27 Licensing software for use on cloud services does not present any particular challenges for research codes, many of which are open-source. The range of licence terms applied by different vendors of closed software preclude general advice, but it is clear that as Cloud Computing becomes more established within research, licences will be adapted to refer specifically to cloud services. It is also possible that software vendors will adapt by offering their software as a service.

Provision and negotiation

- 28 Given that it is unlikely that major cloud vendors will wish to negotiate contract terms with individual academics, there are opportunities for institutional or national engagement, but it is not possible to make overall recommendations about the 'best' approach. This will vary according to the specific research task, the funding available, institutional strategies and policies, and the specific skills of the researcher.
- 29 We recommend against attempting to develop a production UK academic cloud. Such a service would have a real risk of becoming a system that has the appearance of a cloud, some of the functionality of a cloud, the same interfaces, but not the capacity or capability of a true cloud. It is likely to remain the preserve of those who are interested by the technology, rather than those who want to use the technology to meet their research requirements. Cloud Computing is made efficient and effective by its very large scale. This scale permits the flexible and elastic nature of the services that are provided. A UK academic cloud would not have the scale to realise the key benefits of Cloud Computing, yet would still accrue most of the disbenefits. Small-scale clouds for development and testing could be provided locally, but the economics compared to buying capacity from commercial suppliers would need to be considered carefully.

Recommendations

30 The following recommendations are made throughout the document:

- **Recommendation 1:** any organisation considering adopting any cloud services for mission-critical applications, or for processing personal or otherwise sensitive information, should obtain specialist legal advice regarding their contracts and SLAs.
- **Recommendation 2:** JISC should investigate the issues surrounding the management and preservation of research data in cloud systems, and produce guidance aimed at researchers. This should support risk assessment and management, and should not design or develop technical solutions.
- **Recommendation 3:** JISC should investigate mechanisms for national engagement, negotiation and procurement of cloud services, primarily with AWS, Google and MS, but allowing for the engagement of smaller, niche providers.
- **Recommendation 4:** The NGS, and its funders, should consider whether there is a role for that organisation in supporting the development of virtual machine images for common research codes, to allow users to deploy them easily within commercial and private clouds. This may include liaising with or funding the developers or maintainers of the codes.
- **Recommendation 5:** unless backed by clear evidence of demand, and a robust and revenue-neutral business case, JISC should not support the development of a production UK academic research cloud.

This page is intentionally blank

Document history

Version	Date	Description of Revision
0.1	20 April 2010	Initial drafts
0.2	21 April 2010	Combined inputs from multiple authors, added synthesis
0.3	22 April 2010	Version for internal review
0.4	24 April 2010	Version for internal review
0.5	26 April 2010	Version for client review
0.6	7 May 2010	Version including client comments
1.0	7 June 2010	Release version

This page is intentionally blank

List of contents

Executive summary	i
Document history	vii
List of contents	ix
List of abbreviations	xi
1 Introduction	1
1.1 General	1
1.2 Objectives	1
1.3 Scope	1
1.4 Approach	1
1.5 Interaction with Technical Review of Cloud Computing project	2
1.6 <i>Caveat lector</i>	2
1.7 Overview of this report	3
2 Background	5
2.1 Introduction	5
2.2 Infrastructure provision and availability	5
2.3 The research computing 'stack'	7
2.4 Cloud Computing	9
3 Drivers & barriers	11
3.1 Introduction	11
3.2 Political	11
3.3 Economic	12
3.4 Societal	15
3.5 Technological	18
3.6 Legal	20
3.7 Environmental	24
4 Research use cases	25
4.1 Introduction	25
4.2 Computation categorisation grid	26
4.3 Example computational use cases	26
4.4 Research Use Case Scenarios and possible opportunities	34
5 Current cloud offerings	41
5.1 Introduction	41
5.2 Datacentre as a Service (DaaS) providers	41
5.3 Data/Storage as a Service Providers	42
5.4 Amazon Web Services: IaaS	42
5.5 PaaS Providers	43
5.6 Derivative Services	44
5.7 Private Clouds	44
5.8 Contracts and SLAs	46
6 Analysis and conclusions	51
6.1 Introduction	51

6.2	Current situation	51
6.3	Outlook	54
7	Recommendations	57
A	Interviews conducted	59

List of abbreviations

AMI	Amazon Machine Image
API	Application Programme Interface
AWS	Amazon Web Services
DaaS	Datacentre as a Service
DPA	Data Protection Act
EBS	Amazon Elastic Block Store
EC2	Amazon Elastic Compute Cloud
EGI	European Grid Initiative
ENISA	European Network and Information Security Agency
fEC	Full Economic Cost
FoI	Freedom of Information
FTE	Full Time Equivalent
GAE	Google App Engine
HE	Higher Education
HECToR	High-End Computing Terascale Resource ²
HEFCE	Higher Education Funding Council for England
HEI	Higher Education Institution
HPC	High Performance Computing
HTC	High Throughput Computing
IA	Information Assurance
IaaS	Infrastructure as a Service
JISC	Joint Information Systems Committee
NGS	National Grid Service
NIST	National Institute of Standards and Technology
PaaS	Platform as a Service
PAYG	Pay As You Go
PRACE	Partnership for Advanced Computing in Europe
QR	Quality Related [Funding]
RUCS	Research Use Case Scenario
S3	Amazon Simple Storage Service
SaaS	Software as a Service
SAN	Storage Area Network
SDK	Software Development Kit
SLA	Service Level Agreement
VM	Virtual Machine

² HECToR is the UK's largest and most powerful supercomputer. UK researchers can gain access to HECToR by applying through the Research Councils on one of several routes, depending on the details of their requirements.

VPN Virtual Private Network

1 Introduction

1.1 General

1.1.1 Curtis+Cartwright Consulting Ltd, working with the University of Surrey and Professor Charles Oppenheim have been commissioned by JISC to investigate Cloud Computing for Research. This version of the document (1.0) is issued for review by JISC.

1.2 Objectives

1.2.1 The objectives of the project are to:

- document use cases for Cloud Computing in research for data storage and computing;
- develop guidance on the governance, legal and economic issues around using cloud services for storage and computing in academic research;
- make recommendations to JISC on possible further work in the area for data storage and computing.

Timeline

1.2.2 The project ran from November 2009 to April 2010. Interviews were conducted in the period December 2009 – March 2010.

1.3 Scope

1.3.1 The scope of this report is to consider Cloud Computing for research in the areas of compute and storage. Infrastructure as a Service, and Platform as a Service are in scope, while Software as a Service is not.

1.4 Approach

1.4.1 Existing guidance on Cloud Computing for data storage and computing was identified and analysed, considering potential differences between UK HE research use of cloud and that of other sectors/organisations. We have built on readily available conceptual frameworks such as the National Institute of Standards and Technology's (NIST) service and deployment models and the white paper from the Cloud Computing Use Case Discussion Group³ in order to describe a broad range of use cases.

1.4.2 A large number of in-depth interviews were conducted with institutional computing services, research computing services, research council representatives, researchers, and specialist technical staff (See Annex A) to understand more fully the drivers and barriers that affect the use (and potential use) of Cloud Computing in research, and to develop further use cases and case studies.

1.4.3 This document (and the accompanying briefing paper) are intended to provide decision support, and not a single recommendation: critical thinking is still required from researchers and institutions as to what data storage or compute solution is most appropriate given

³ <<http://groups.google.com/group/cloud-computing-use-cases>> [accessed 23 April 2010].

functional requirements, budget, security, reliability, trust, *etc* as well as the cloud services currently on offer.

1.5 Interaction with Technical Review of Cloud Computing project

- 1.5.1 This project, '*Cloud Computing for Research*' has a sister project called '*Using Cloud for Research: A Technical Review*'. The two projects are designed so that broad use cases, economic considerations, wider societal barriers and enablers and legal issues fall within the scope of this work (Cloud Computing for Research). The technical review covers the technological capabilities of current cloud provision, details of specific hypervisors *etc*, and performance of specific codes.
- 1.5.2 However, there is inevitably some overlap between the projects. This happens particularly when considering specific examples of researchers using clouds for research and in discussing general use cases. Whilst we do offer our opinion of the applicability of Cloud Computing to *general classes* of research, as a general rule we have avoided discussion of the performance of *specific applications*.
- 1.5.3 Should the reader require more detail on (virtual) hardware, application benchmarking, and the relative performance of the offering from different cloud vendors, then they are referred to the Technical Review. In addition, a range of online resources on this topic are readily available⁴ and a brief outline of the academic research currently being done in this area is included at Annex F.

1.6 Caveat lector

- 1.6.1 There is currently an extremely rapid pace of change and innovation across the range of activities considered Cloud Computing. This covers not just the details of costs of services from specific providers, but also the range of services offered and the available tools for configuring and controlling cloud resources.
- 1.6.2 We present here a snapshot of the situation in April 2010. All readers are strongly urged to keep in mind the fast changing nature of the field: our descriptions of services will be out of date within months from publication, and entirely new ideas, technologies, business models, providers, strategies and risks will continue to appear.

⁴ For example, www.genomeweb.com/sites/default/files/walker.pdf, www.cs.st-andrews.ac.uk/files/PerformanceComparison.pdf and www.cs.st-andrews.ac.uk/files/PlatformComparison.pdf.

1.7 Overview of this report

1.7.1 The remainder of this report is set out as follows:

Section	Contents	Target audience(s)
Section 2	Describes the research computing environment, introduces the concepts of Cloud Computing, and briefly outlines the current provision of cloud services. This section provides general background information, and defines some terms which appear throughout this report.	<ul style="list-style-type: none"> – All readers
Section 3	Sets out a range of drivers and barriers that will influence the applicability and uptake of Cloud Computing within research in the UK. Drawn from a large number of in-depth interviews with institutional computing services, research computing services, research council representatives, researchers, and specialist technical staff, this section describes the range of opinions and concerns expressed by stakeholders.	<ul style="list-style-type: none"> – Those making strategic decisions regarding policies or investments. – Vendors or potential vendors of cloud services
Section 4	Describes the range of use cases for Cloud Computing in research, including case studies of active research, and proposes use cases where cloud approaches may prove particularly beneficial in future. Detailed case studies are included in Annex C.	<ul style="list-style-type: none"> – Active researchers – Research managers
Section 5	Outlines the services offered by the market leaders. Analyses the contracts and SLAs offered by the three leading providers – Amazon, Google and Microsoft. Further information, including information on other providers is included in Annex E.	<ul style="list-style-type: none"> – Active researchers – Research managers – Those making strategic decisions regarding policies or investments – Institutional risk managers, including contracts and compliance officers
Section 6	Synthesises the range of information collected during this study, and presents the overall view of current and possible future uses of Cloud Computing in research.	<ul style="list-style-type: none"> – All readers
Section 7	Lists recommendations from this project.	<ul style="list-style-type: none"> – All readers
Annex A	Lists the interviews conducted for this review.	

1.7.2 The following annexes are contained within the accompanying annexes document:

Section	Contents
Annex B	presents an expanded Cloud Computing taxonomy based on NIST and Gartner definitions, and the Cloud Computing Use Case Discussion Group White Paper.
Annex C	presents the expanded case studies of researchers using clouds and the responses to a questionnaire of overseas users of cloud
Annex D	provides details on the researchers who responded to our survey
Annex E	provides further details on current cloud providers
Annex F	lists some of the research, and research areas, currently being undertaken into Cloud Computing
Annex G	Provides some examples of datasets hosted by Google

This page is intentionally blank

2 Background

Contents	Target audience(s)
describes the research computing environment, introduces the concepts of Cloud Computing, and briefly outlines the current provision of cloud services. This section provides general background information, and defines some terms which appear throughout this report.	– All readers

2.1 Introduction

2.1.1 This section briefly answers the question “what is research computing?” and describes the research computing environment, introduces the concepts of Cloud Computing, and briefly outlines the current provision of cloud services. Information has been gathered from wide-ranging interviews with institutional computing services, research computing services, research council representatives, researchers, and specialist technical staff.

2.1.2 More detail on specific cloud providers is given in Section 5.

What is research computing?

2.1.3 For the purposes of this work, the project team has considered research computing to be computing focused on addressing applied research problems rather than the area of computer science. The term ‘research computing’ encompasses the entire ecosystem of infrastructure, researchers, research computing services and IT services, institutions and research funders.

2.1.4 Two broad areas of computing emerge as traditionally making use of different infrastructure:

- **High Performance Computing (HPC):** is focused on the cluster and supercomputer. These machines are designed to perform the maximum number of operations per second, and make use of special architectures to achieve this goal. A key characteristic HPC machines share is a low-latency interconnect, such as InfinBand, which makes it possible to share data very rapidly between large numbers of processors working on the same problem.
- **High Throughput Computing (HTC):** the defining principle of HTC is that, for many researchers commodity CPUs, memory and networking are sufficient. In order to increase research productivity the focus should then be on maximising the throughput over a long period – in other words ‘sweating the assets’. Access to a large pool of machines, each one modest in its capabilities, can allow researchers to perform tasks such as parameter sweeps much more rapidly than being limited to a single workstation.

2.2 Infrastructure provision and availability

2.2.1 Cloud Computing offers a new route to accessing computational infrastructure. It is therefore useful to consider at the outset the current range of infrastructure that is available to researchers.

Desktops and workstations

- 2.2.2 Desktop and workstations (usually meaning a more powerful desktop) are the lynchpin of research computing – the majority of a researcher’s day is spent in front of their workstation. They are usually based on the Linux or Windows platforms and run the traditional productivity software (e-mail, word-processing and spreadsheets *etc*), have internet access, and also perform calculations and analyses. Internet access is very important since desktops act as the gateway to accessing remote resources, such as institutional clusters, through protocols such as SSH or Grid certification.
- 2.2.3 The low cost of large capacity hard-drives means that researchers can have a large amount of cheap storage in their workstation. In terms of cost the more robust storage options offered by, say, central IT services can appear unfavourable by comparison. However, over the course of a 3-year funding agreement the prudent researcher should plan for an inevitable hard-drive failure, and make suitable backup provision.
- 2.2.4 The growing power of individual workstations has increased their importance as a computational resource. This power can be harnessed by linking desktops together into grids, in which groups of desktops process tasks allocated by a workload management system such as Condor.⁵ These grids are a key resource of HTC.

Local cluster

- 2.2.5 Research groups may sometimes own and operate their own ‘clusters’. They can be as simple as a collection of desktops networked together to form a cluster, although more advanced hardware is also available. These clusters are often housed in areas not intended to be used as machine rooms – the so-called ‘broom-cupboard cluster’.
- 2.2.6 It is growing more common, however, for groups purchasing their own computer to do so as additional ‘nodes’ that are added to larger institutional resources. The research computing service then takes on responsibility for housing and maintaining the nodes, to which the research groups gets sole, or preferential, access.

Departmental cluster

- 2.2.7 Departments, Institutes, Schools or Faculties, may provide a cluster for the benefit of the whole department. The cluster is shared between research groups in the department, who may each contribute to its cost.

Institutional cluster/HPC

- 2.2.8 Often an institution will host a large-scale computational facility available to all researchers within the institution. These facilities are operated and maintained by central IT services or dedicated research computing centres. Institutional computing facilities can be very powerful; the University of Southampton’s new Iridis 3 machine ranks highly in the most recent Top500 list.⁶

Shared services

- 2.2.9 By going outside the bounds of their host institution a researcher can also access other resources, often at even larger scale. Some of these resources are:

⁵ <www.cs.wisc.edu/condor/> [accessed 12 April]

⁶ <www.top500.org/list/2009/11/100> [accessed 12 April 2010]

- **NGS:**⁷ the NGS (National Grid Service) aims to enable coherent electronic access to distributed computing resources for all UK researchers. Accessing resources is then completely independent from researcher or resource location. Institutions volunteer to make their resources available to NGS users. The NGS does not currently charge for compute time, but volunteering institutions are free to do so. The NGS ensures compatibility of standards between resources, and defines a list of applications installed on all NGS nodes.
- **HECToR:**⁸ the High-End Computing Terascale Resource is the UK's largest and most powerful supercomputer.⁶ It is funded by the Research Councils and managed by the EPSRC on behalf of the others. UK researchers can gain access to HECToR by applying through the Research Councils on one of several routes, depending on the details of their requirements.
- **PRACE:**⁹ the Partnership for Advanced Computing in Europe aims to create a pan-European HPC service of several petascale facilities supported by national supercomputing centres. It is currently at the prototype stage.¹⁰

2.3 The research computing 'stack'

2.3.1 It is important to note that research computing is much broader than the raw hardware available. There is an entire multi-levelled ecosystem sitting atop the 'bare metal'. The infrastructure described in the previous section forms the bottom layer of what we have termed the 'research computing stack'.

Infrastructure

2.3.2 The lowest level of the stack is infrastructure, which is the physical hardware that powers and makes possible research computing. The upper levels of the research computing stack, or parts of it, exists on each of the different infrastructures described at section 2.2. Infrastructure is traditionally split into:

- **networking:** the connections between machines along which data flows;
- **compute:** the 'guts' of a computer – CPU, RAM *etc*;
- **storage:** hard-drives, SANs, and other storage devices.

Virtual machines (VMs)

2.3.3 The virtualised layer is not usually present in research computing infrastructure, but is revolutionising enterprise systems and is a key enabling technology in Cloud Computing. The principle is that any given hardware can be virtualised into a number of *virtual machines*. The hardware is thus abstracted from the user, who only sees the virtual machine. These machines exist and operate independently from each other.

2.3.4 Virtual machines may have a number of advantages, including reduction of costs by allowing higher utilisation of existing physical servers. However, the additional layers of software carry an overhead in that they can reduce system performance.

⁷ <www.ngs.ac.uk/> [accessed 12 April 2010]

⁸ <www.hector.ac.uk/> [accessed 12 April 2010]

⁹ <www.prace-project.eu/> [accessed 12 April 2010]

¹⁰ <www.prace-project.eu/press-corner> [accessed 12 April 2010]

Platform

- 2.3.5 A 'platform' is a self-consistent computing environment that sits between the user and the hardware of the system. Operating systems such as Windows and the Linux variants are platforms. In Cloud Computing the concept of Platform-as-a-Service is one of the main business models.
- 2.3.6 Key components of the platform are:
- the **development environment** which is a set of applications that allow software developers to write code for that platform. That code will then run on any machine running that platform.
 - and **libraries** which are self-contained packages of code that are used by other software. Rather than code from scratch each time, developers will use libraries for many common tasks.

Applications

- 2.3.7 Applications are pieces of software with which the user directly interacts, and uses to do everything from checking their email to submitting jobs to a compute cluster. In addition to the self-written code and scripts used by researchers, there are two classes of application particularly relevant to the Cloud Computing debate:
- **SaaS/Portals:** Software-as-a-Service and portals are both methods of accessing an application remotely; the application is not installed on the researcher's personal machine but is instead accessed through a browser-type interface. Both the application and infrastructure it runs on are remote from the user.
 - **Managed scientific code:** Many researchers use applications written by, and for, scientists. This software is often available free-of-charge under an Open Source license. The code is kept up-to-date and improved periodically; each release being a new 'version'. It is the user's, or systems administrator's, responsibility to ensure that the latest version is installed on their system. Examples of managed scientific code include the numerous applications covered by the aegis of the Collaborative Computational Projects Group.¹¹

The cloud stack

- 2.3.8 The Use Cases for Cloud Computing White Paper¹² presents a "taxonomy" of the Cloud, showing how the various elements of the stack presented above are an inherent part of Cloud in general, and how these are complemented by capabilities such as Security and System Management.
- 2.3.9 The Cloud offerings of Infrastructure, Platform and Software may be supported by virtualised systems, and accompanied by virtual images that contain the operating system and platform elements or required applications¹³. The virtual images run in VMs; there are various ways in which the virtual images can be created and formats in which they can exist.

¹¹ <www.ccp.ac.uk/> [accessed 12 April 2010]

¹² Available via <http://cloud-computing-use-cases.googlegroups.com/web/Cloud_Computing_Use_Cases_Whitepaper-2_0.pdf> [Accessed 21 April 2010]

¹³ See, for example, <<http://virtualboximages.com/>> [accessed 12 April 2010]. A typical example is a virtual image that contains a LAMP, or other *AMP, stack for providing a web server. See: <[http://en.wikipedia.org/wiki/LAMP_\(software_bundle\)](http://en.wikipedia.org/wiki/LAMP_(software_bundle))> [accessed 12 April 2010] for further information.

- 2.3.10 The Cloud Stack also helps to classify the variety of research undertaken, in relation to Cloud Computing. See Annex F for examples of research published in relation to these elements.

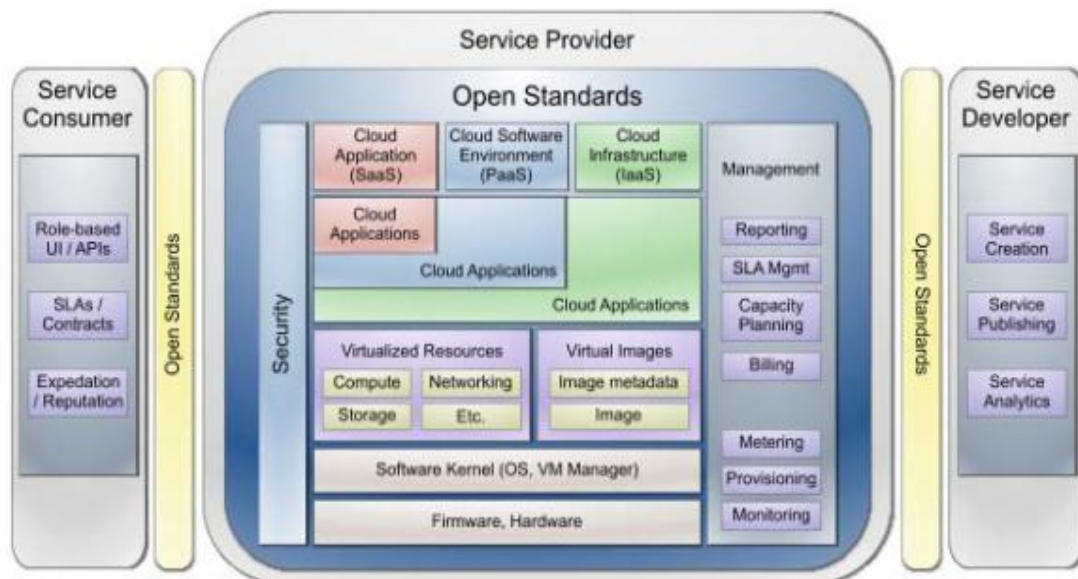


Figure 2-1: the cloud stack. From Cloud Computing Use Cases White Paper, version 2.0

2.4 Cloud Computing

- 2.4.1 At the time of writing, “cloud” as a term is being applied to a broad range of computing activities, some of which are novel, and some (*eg* webmail) which are already well-established.
- 2.4.2 Cloud Computing variously combines considerations of hardware, software, services, programming environments, resource management, billing services, and legal, regulatory and ethical governance.
- 2.4.3 Various definitions and characteristics of Cloud Computing have been put forward; the NIST definitions¹⁴ provide reasonable coverage, though some of the distinctions may not be as granular as might be desired.
- 2.4.4 NIST defines Cloud Computing as:
- “a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (*eg*, networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. This cloud model promotes availability and is composed of five essential characteristics, three service models, and four deployment models”.
- 2.4.5 The characteristics, service models and deployment models used by many to characterise Cloud Computing are elaborated further in Annex B.

14

<<http://csrc.nist.gov/groups/SNS/cloud-computing/cloud-def-v15.doc>> [accessed 21 April 2010]

This page is intentionally blank

3 Drivers & barriers

Contents	Target audience(s)
Sets out a range of drivers and barriers that will influence the applicability and uptake of Cloud Computing within research in the UK. Drawn from a large number of in-depth interviews with institutional computing services, research computing services, research council representatives, researchers, and specialist technical staff, this section describes the range of opinions and concerns expressed by stakeholders.	<ul style="list-style-type: none">– Those making strategic decisions regarding policies or investments.– Vendors or potential vendors of cloud services

3.1 Introduction

- 3.1.1 This section considers a range of drivers and barriers that will influence the applicability and uptake of Cloud Computing within research in the UK. It is structured according to the PESTLE format – Political, Economic, Societal, Technological, Legal and Environmental. These drivers and barriers have been synthesised from the discussions and interviews with the stakeholders and persons listed in Annex A. These include, amongst others, representatives from the research councils, institutional research computing and IT services directors, and current researchers.
- 3.1.2 This section is primarily focused on the *current situation* within research computing, and on typical research computing tasks, which currently demand dedicated HPC or HTC clusters. Many of the potential benefits of cloud approaches for research computing are in the opportunity to address new kinds of problems, or to meet requirements which are currently latent – ideas which researchers may have, but have never been expressed because the infrastructure is not there to support them. This section does not address these potential future benefits directly. See section 4.4 for analysis of these new, disruptive, opportunities.

3.2 Political

Institutional strategy

- 3.2.1 An institutional-level decision is whether to invest in research computing facilities. A range of options is available, including broadly:
- Invest in an institutional HPC and/or HTC facility, and mandate its use;
 - Provide organisational structures to deliver and support central resources which are funded locally;
 - Allow departments or individual researchers to provide and maintain their own facilities.
- 3.2.2 The approach taken will depend on the degree to which the organisation intends to develop research computing as a strength, and the level and diversity of use of extant resources. Institutions will need to balance the overheads in providing compute facilities with the benefits of owning their own equipment.

National strategy

- 3.2.3 Presently, national policy regarding research computing is driven through the research councils, each of which has a range of specific research computing requirements and has developed their own approaches to the provision of HPC. Shared facilities such as HECToR

are a clear example of attempts to produce national-scale economies in the procurement of HPC.

- 3.2.4 Consideration of so-called “medium performance computing” is less well developed. These tasks do not demand the exotic architectures of dedicated supercomputers, but benefit from large-scale parallel processing of the type that is easily provided by cluster resources. Almost all research-focused institutions support at least one cluster, and most have several. However, the current strategic environment does not provide a drive toward rationalisation or sharing of these services.
- 3.2.5 Broader questions of international collaboration and competition are also relevant here; how can the UK compete with the vast investments in HPC in the US? Should the UK buy into PRACE?¹⁵ What will be the role of EGI¹⁶ in large-scale computing, and how will this affect the NGS?

The funding of research computing

- 3.2.6 See also economic drivers and barriers below.
- 3.2.7 The dual support model of research funding leads to a range of funding sources for institutional compute infrastructures. Research councils are content to provide some funding for compute resources within project funding, subject to the research case being made adequately within the proposal. The councils typically see larger investments¹⁷ as infrastructure which should be funded by the host institution – and this typically means the institution committing Funding Councils’ Quality-Related (QR) block grant funding to the investment, either at a departmental level or more broadly (see *Institutional strategy* above). Strategic investments by research councils are typically considered in a different manner from typical research proposals.
- 3.2.8 The research councils are at an early stage in forming views on the impact of Cloud Computing on research computing. Policies are nascent, and they are not being pressured by their stakeholders to consider cloud approaches explicitly. Research council policies that drove uptake of shared facilities or cloud would be strong, and would likely have knock-on effects in reducing the number of smaller clusters in use around the country.
- 3.2.9 Moving from an infrastructural investment model to an on-demand pay as you go (PAYG) model could release QR funds for other purposes, leaving the research councils to fund cloud provision for specific projects as a directly allocated cost.

3.3 Economic

Hidden costs in local provision

- 3.3.1 In many institutions, the full cost of research computing is not exposed to the users. By historical legacy, or by choice, grant holders typically do not have any directly allocated charges for the use of research computing facilities. The key elements which are usually “hidden” in overheads are power, cooling, and support staff. These costs are typically distributed amongst a larger group of grant holders – either amongst a department, a research grouping, or even across the entire institution.

¹⁵ The Partnership for Advanced Computing in Europe - <<http://www.prace-project.eu>> [accessed 12 March 2010].

¹⁶ The European Grid Initiative - <<http://web.eu-egi.eu>> [accessed 12 March 2010].

¹⁷ What is considered a large investment varies depending on the council.

- 3.3.2 With so many of the costs of institutional provision being hidden from the users of the service, cloud services – which expose the entire cost of provision – often seem expensive. Unless an organisation has fully identified the costs of local provision, and understood how those costs would be changed by adoption of cloud systems, it is not possible to judge which approach is less expensive overall.

Cloud Computing is not yet a major concern for research councils

- 3.3.3 The research councils are agnostic to the approach taken to providing computational resources for projects they fund. When proposals are reviewed, the case for any investment in new facilities is considered – but there is no requirement to compare the selected approach with other approaches.
- 3.3.4 For example, a proposal may clearly show how a particular computational problem will require a new workstation computer. There is no need to consider whether this computational effort could be provided more efficiently or effectively using an existing institutional cluster, a national service, or a cloud solution. Although the solution for that particular proposal may be effective, this does not lead to rational investment across the range of projects and institutions funded.
- 3.3.5 At present the position of the research councils regarding cloud can be best summarised as “curious.” There is some interest amongst some computational researchers, but the councils do not have clear policies on Cloud Computing. They do not report strong demand from their grant recipients (or potential grant recipients) for them to take a position regarding cloud.

Double-charging research councils

- 3.3.6 There is a risk that directly allocating the costs for computing could lead to accusations of “double charging” for research computing. This would be the case where an overhead charge was present within the full Economic Cost (fEC) for compute infrastructure, and a directly allocated charge was also applied.
- 3.3.7 Several institutions raised this as a concern, considering that the councils might feel that they were funding the institution to deliver infrastructure as an overhead charge, but also paying for usage. Conversely, several other institutions have established clear mechanisms for managing directly allocated computing costs alongside overheads, for example, when a directly allocated charge can be used to provide preferential access to a shared institutional HPC facility, or to provide extra nodes for the facility. The research councils have expressed views that this “double charging” approach is not problematic, as long as it is applied transparently.
- 3.3.8 There is a broader question regarding whether it is beneficial to expose the costs of research computing, or to roll them up as an overhead charge. Calculating the full cost of the provision of research computing is very difficult, and few institutions have attempted to do this comprehensively. This can lead to a situation where an agreed overall institutional cost for computing services (*eg* a charge per node hour) appears expensive to academics, who have had the true costs of this infrastructure hidden from them previously (see above). This is particularly important when submitting research funding proposals, where the proposal reviewers will have their own ideas of the cost of provision of research computing, which may not be particularly well aligned with the true cost.

Trading capital expenditure for operational expenditure

- 3.3.9 The shift from a capital expenditure to ongoing, and probably variable, revenue expenditure can cause an institution to be exposed to a degree of risk. This is particularly the issue where research requirements may not be well understood – it may not be possible to tell at the outset of a project how much compute and storage requirement will be required. In the short term, moving traditional research computing tasks to cloud provision of computing will involve significant uncertainties; the change of platform and provision will affect efficiency, and hence the amount of resource required, and hence the cost. However, before the tasks are run it is impossible to know what the performance implications are, and of course running benchmarks on cloud systems has a cost implication itself.

Balancing costs and efficiencies

- 3.3.10 Cloud Computing has the potential to provide effectively limitless compute and storage capacity to researchers. The ready availability of capacity may lead to changes in the way that applications are designed – decisions about whether it is worth spending significant time and effort on optimising an algorithm *versus* just applying more resources to the computational task may be changed in a cloud environment.
- 3.3.11 There are two key factors which will tend toward an emphasis on efficient software design. Firstly, most computational tasks do not scale linearly – the ‘law of diminishing returns’. In fact, some codes may reach a point where adding more nodes to the task actually increases the time taken! Secondly, cloud resources are not free – there is a direct cost in using more resources. This may actually lead to more rational decisions being taken about how to spend effort on research problems – improve the code, or brute-force the problem?

Economic pressures are driving efficiencies

- 3.3.12 There is a significant desire to increase the efficiency of research. At present, this is primarily driven by the institutions (and their component parts) who aim to increase their competitiveness. Although these efficiencies are primarily local at present – a machine room shared by several departments, a shared systems administrator, a shared HPC facility *etc* – there is a developing awareness that larger-scale efficiencies might be possible. At present, this is typically viewed as the opportunity to collaborate, or to create shared services (such as the proposed Shared HE Datacentre)
- 3.3.13 It seems likely that research councils will come under increasing pressure to ensure efficiency of research as well as effectiveness. The current economic situation is likely to accelerate the current rationalisation and reduction of investment.

You can't employ half a systems administrator

- 3.3.14 Personnel typically come in discrete units! Few research grants are of a scale that will support a full-time system administrator, so these roles are often shared between grants, or shared with research responsibility. A typical argument in favour of the adoption of cloud approaches is that by doing so, it is possible to avoid the overhead involved in maintaining local infrastructure, which may be hidden (*eg* a postdoc who, as well as undertaking their funded research, has taken on the management of the research group's cluster). However, the number of instances where an individual maintains compute infrastructure unfunded is probably limited.

- 3.3.15 Although having split roles may not necessarily be the most effective way to deliver the systems administrator function, it may have other benefits. If a research grant can cover, say 0.25 FTE of a system administrator, along with other sources of funding this may allow a group leader to keep on a valued postdoc who could not otherwise be employed. This may allow skills to be retained within a research group, which is often a key goal for group leaders.
- 3.3.16 Many institutions have systems administrators employed centrally, or on a shared basis between research groups. Whether these individuals are funded through overheads, or through the combination of contributions from a range of grants varies.

Benefits and costs are misaligned

- 3.3.17 The key benefits of Cloud Computing are held to be:
- On-demand computing power and pay-as-you-go charging, which avoids over-buying resources;
 - Greater efficiency of provision as economies of scale are realised;
 - Flexibility, with (as on Amazon's EC2 service as an example) the ability to create new compute resources to experiment with;
 - Reduction or removal of capital investment required.
- 3.3.18 Considering the typical research computing use cases, many of these potential benefits of adopting Cloud Computing accrue at an institutional or national level rather than to the researchers themselves. It may indeed be a disbenefit to adopt cloud services (for the reasons outlined elsewhere within this section). This misalignment of benefits and costs is likely to preclude extensive natural adoption of cloud technologies.

3.4 Societal

The joys of ownership

- 3.4.1 There are a number of issues relating to perceptions and feelings about ownership of computing resources (at the institutional and individual scale). These issues are explored in more detail in the sections below.

Prestige

- 3.4.2 Put quite simply, having a large scale 'world-class' computing resource onsite confers prestige upon an institution or research computing centre.¹⁸ It is seen as recognition of the institution's commitment to the disciplines of research computing, its expertise in the field, and its success. Ownership of prestigious facilities is also important in the competition to attract the most talented and productive researchers.

¹⁸

Notwithstanding the prestige of having significant on-site compute resources, many experienced researchers and computing staff find it amusing that the machine room is often a featured highlight of tours for visiting dignitaries and prospective staff and students. Whilst the noisy air-conditioning and intimidating server-racks are undoubtedly impressive for the uninitiated, the vast majority of users will never need to set foot in the room again.

Control and academic freedom

- 3.4.3 Control of computational hardware and resources can be tied to the elusive concept of academic freedom so craved by researchers of all kinds. The idea of 'control' is manifested in the authority and ability conferred by ownership to choose how a resource is managed and configured. The benefits of direct control of resources may actually be greatest for the small-scale resources directly maintained by individual research groups. However, is ownership necessary to have control? IaaS cloud offerings feature root-access to the virtual machines, which are requested by the users. This is a level of control not enjoyed by some researchers for their own workstations.
- 3.4.4 A slightly different form of control is centred on the availability of the support staff and systems administrators. Many researchers feel that having the systems administrators to hand increases their accountability, perhaps best summarised as 'the ability to go and shout at someone': if a problem should occur, then having both the hardware and systems administrator available locally is perceived to reduce the time taken to restore normal operation. But is this *only* a perception? Institutional systems administrators are subject to the same institutional closures, over the Christmas period and bank holidays, that affect all institution staff. There is unlikely to be a 24/7 helpdesk for research computing systems. They also have more demands on their time than merely monitoring the performance of, say, the institutional HPC cluster. Problems can potentially go unnoticed until a user draws attention to them.
- 3.4.5 Commercially provided cloud resources are based in data-centres that are, or should be, monitored constantly, 24 hours a day, 365 days a year. Indeed, dynamic resource provision coupled with data redundancy means that, in theory at least, hardware failures should go unnoticed by the end-user. This concept, that computing resources are not tied to specific hardware, is at the heart of the Cloud Computing ideal.

Accessibility

- 3.4.6 'Available anywhere, anytime' is a well-rehearsed phrase in the hype that surrounds Cloud Computing, but it is a real advantage if a researcher is working away from their home institution or wishes to, say, make a data-set widely available.
- 3.4.7 Access policies to institutional computational resources can vary across institutions, and it is not unknown for access from outside the institutional firewall to be prohibited. Cloud services, by their very nature, have to be internet-accessible, so they are reachable from anywhere with an internet connection.

Researcher training and development

- 3.4.8 A vital function of university research is the development of the skills and experience of the researchers themselves. PhD students especially may start their postgraduate career from a relatively low knowledge base. A truism that many PhD students experience is that it is not until halfway through their second year that they learn enough to know that "everything they've done so far is no good". Their supervisors of course knew this all along. Providing an environment in which researchers can develop their abilities is a function filled by local resources in most institutions.

Scaling-up calculations

- 3.4.9 Many forms of research computing follow a pattern of using progressively larger resources. For example, in the field of molecular simulation the simulations can be designed and built on desktop computers, then local clusters might be used for testing and investigating the systems and hypotheses. Large amounts of compute time, either on institutional or national-scale facilities, will then be used to accumulate long trajectories.
- 3.4.10 The kind of development route described above utilises different scales of hardware, possibly administered by separate organisations, which then necessitates multiple user accounts (the work of the NGS in the problem area of user authentication notwithstanding). In a cloud environment, however, the issue of scale is trivial; resources of the same type as are currently being used become available on demand.

The effect of pay-as-you-go on innovation

- 3.4.11 An oft-repeated maxim in the business world is that failure is integral to the process of innovation, and that raising the cost of failure is the same as raising the cost of innovation. Some researchers fear that a PAYG charging-model in which every CPU hour and GB of storage incurs a monetary charge at the time it is consumed will stifle innovation by making both funders and researchers wary of expending resources on risky or highly experimental projects.
- 3.4.12 Interestingly, one of the main selling points of Cloud Computing is that it actually lowers the cost of innovation by eliminating the 'start-up tax' of purchasing hardware (which is wasted capital if the innovation is a failure). To what extent is this view applicable to research computing? Unlike venture start-ups, research does not end if the initial idea fails. Research funders and academics alike both know that the funding proposal can be viewed as a statement of intent and not necessarily of what will definitely be achieved. Researchers work on new approaches to their area of study if their first, second and third attempts prove fruitless.
- 3.4.13 However, it is conceivable that 'spontaneous innovation' – the leftfield idea born out of a desire to try something new and exciting – may be curtailed by a desire to save budget for the main strand of research or for the 'safe bets' where researchers know that publishable results can be generated.

Learning curves

- 3.4.14 There is undeniably a learning curve involved in using any unfamiliar computational resource. Researchers who already have relevant experience to draw on often find it easy or trivial to move to a new platform or hardware. However, for those researchers new to research computing and for non-traditional users, especially, the barrier to getting started can be highly off-putting.¹⁹
- 3.4.15 It is easy to forget that the command-line driven interface to most computational platforms can be a daunting prospect for those unfamiliar with it, but does Cloud Computing offer an easier option? Proponents often claim that it does; however, configuring and operating an IaaS offering does require some technical knowledge, and the command-line is often no less important.

¹⁹ Relevant JISC work exploring the barriers and enablers for non-traditional users of research computing includes the ENGAGE, eUptake, eUIS, projects (all found at <<http://engage.ac.uk/engage>> [accessed 23 April 2010].

- 3.4.16 The case studies in Annex C contain reference to the length of time it took computationally experienced researchers to get started on a cloud platform. Most of the interviewees described the process in terms of being 'easy' but acknowledged that there was a period of time spent reading online resources *etc* and that they experienced some false starts. Spending a few weeks getting up-to-speed is perhaps less of a barrier for these computational researchers who, by nature and inclination, expect to encounter this type of task. For non-traditional users of research computing services the investment in time and energy may be significant indeed.

Cloud is already here

- 3.4.17 There are two major ways in which Cloud Computing is 'already here' for research computing. The first is that since Cloud Computing is already affecting the commercial sector, so it then becomes important for the graduates and postgraduates leaving academia for industry to have relevant experience that enhances their portfolio of skills.
- 3.4.18 The second is that researchers are already using Cloud Computing. Indeed, Section 4 (expanded at Annex D) describes use cases gathered from researchers performing research on a cloud platform. In providing educational grants directly to researchers, Amazon has effectively performed an 'end-run' around the Research Councils and institutions. Whilst these bodies are still ruminating on policy there is a growing band of researchers able to understand firsthand the benefits and disbenefits of Cloud Computing for research. The level of individual grants awarded is small (~US\$5,000), but this is enough to both get started and perform some meaningful research for many workers. The major question mark is over how long Amazon will continue to act as a benevolent funder of research.

XaaS is unknown in research computing

- 3.4.19 Research computing generally conforms to a do-it-yourself model in which the onus is on the individual researcher to locate and learn how to use relevant codes, resources and infrastructure. The more commercially orientated paradigms of SaaS and PaaS (less so for IaaS) offer an environment in which the end-user is relieved of the responsibility of software installation and maintenance. The idea that researchers could simply create and submit jobs through a portal that then handles scheduling and load balancing is alien to many researchers.²⁰
- 3.4.20 In such an environment it becomes much more important to understand the boundary between what is provided as a service and what is left to the user. Service Level Agreements (SLAs) are an important tool in defining the relative responsibilities of provider and user. A discussion of the SLAs offered by cloud providers can be found in Section 5.

3.5 Technological

Support

- 3.5.1 The technical support required by researchers varies greatly according to skills, experience and research needs. However, when making use of institutional facilities a researcher can typically expect to encounter a dedicated systems administrator responsible for the smooth operation of the facility. Processes such as setting up of user accounts and accessing the

²⁰ The NGS, however, does operate a Portal service with is described as an "online repository of pre-configured application templates".

resource, bug reporting, and installation of standard codes, packages and libraries are usually supported as standard.

- 3.5.2 Institutional systems administrators often also provide more specialised help on request with, for example, compilation of code, advice on resource requirements and arranging special access to resources. The background of institutional systems administrators is often research-based, so they are well suited to understanding the special requirements of research.
- 3.5.3 Provision of support for smaller-scale resources, owned by individual groups, is more complicated. Institutional computing centres usually now offer to host, for a price, hardware procured for specific research groups, who will then benefit from the level of support described above.
- 3.5.4 However, when researchers operate their own clusters independently, support is limited to what can be provided by the group's designated systems administrator. Time spent by academics or researchers maintaining and supporting this resource is time taken directly from active research. Despite this limitation, some researchers prefer this arrangement for the increased freedom and control they feel it gives them (see paragraph 3.4.3).
- 3.5.5 Crucial to the discussion of how Cloud Computing could affect research computing is the understanding that the support needs of researchers remain largely the same whether they are using local or remote resources.

Compatibility

- 3.5.6 The vast majority of institutionally owned research computing facilities use Linux-based operating systems, as do most of the workstations that researchers use for their day-to-day tasks. This arrangement provides for ready compatibility and ease of remote access between systems (*eg* a researcher's workstation and departmental compute cluster).
- 3.5.7 There is, however, a sizable group of researchers who use Microsoft-based workstations. Scaling-up, or speeding-up, their computations can be impossible since clusters are nearly all Linux-based. Researchers can find themselves caught in a trap in which they are limited by the speed and power of the machine under their desk. The flexibility of virtualised platforms could make getting access to greater computing power much easier for these researchers.
- 3.5.8 For users who currently depend on Windows workstations and desktops especially, the Microsoft Azure service, or the Windows VMs offered by Amazon Web Services (AWS), might offer a chance to change dramatically the way they do their research.

Characteristics of the research, hardware and application performance

- 3.5.9 Research computing problems (in this case meaning the application of computing to a research area rather than inherent computer science problems) are often described in terms of being HPC or HTC (see paragraph 2.1.4), or compute-intensive *vs.* data intensive. For maximum efficiency, these distinctions require different hardware configurations. Cloud Computing uses virtualisation technology which claims to make the performance of the virtual machine being offered independent of the underlying hardware, but to what extent is this i) true and ii) important for research computing?
- 3.5.10 There is literature available that suggests the performance of applications in the cloud is slower than comparable clusters.²¹ Virtualisation does therefore affect performance,

²¹

<www.genomeweb.com/sites/default/files/walker.pdf> is one example [accessed 19 April 2010].

particularly where communications are concerned. Clouds based on large data centres find it hard to compete with HPC clusters on communication performance. Section 4 contains examples of research computing across a range of computational scales and requirements, which provide real researcher experience and views on this issue.

- 3.5.11 Codes currently being used for research computing may have been written with a specific technical architecture in mind *eg* a low-latency cluster. The act of writing this code to a standard where it has become a useful tool, and perhaps very widely used, can represent a significant sunk investment. In addition to the potential difficulties of porting and compiling codes written for one platform onto a cloud platform, there exists the possibility that the code was written *specifically* for that platform and *optimised* accordingly. Its performance will therefore appear worse on cloud.

3.6 Legal

Data protection

- 3.6.1 The 1998 Data Protection Act applies to all personal data, *ie* data that is about a living identified or identifiable individual, irrespective of where he or she lives or works, that is either managed or is held in the UK. For any Cloud Computing application relevant to a UK-based HEI, the Act will apply because the HEI in question is responsible for the processing (*ie* addition, deletion, editing, manipulation or dissemination) of the personal information. This applies even if the actual processing takes place in another country, or, indeed, in several countries, some of which may or may not be known, as is typical for cloud applications.
- 3.6.2 The Act imposes on the data controller (a legal term which means the HEI) and on any sub-contractor used by the data controller (*ie* the Cloud Computing organisation) certain obligations. It is a breach of the Act if the HEI fails to fulfil its obligations, or if the HEI fails to impose those obligations on its sub-contractors. This applies wherever the sub-contractors are based and whatever legislative environment they happen to work in. The best way to achieve it is to have a clause in the agreement with the supplier that the supplier shall at all times observe and obey the requirements of the Data Protection Act 1998 whilst handling personal data belonging to the HEI. An alternative is for there to be an explicit list of obligations (which happen to be those required by the Act) imposed on the cloud service supplier either in the contract or as a Schedule to that contract.
- 3.6.3 Personal data handled by HEIs in a research context include material on staff, students, research associates, individuals who happen to be the subject of a research project, and individual contractors, suppliers and partners. The data can range from the most innocuous (*eg*, authors' names in a bibliography of a research report, the name of the research associate responsible for particular actions, or the web pages of members of staff) through moderately sensitive (such as e-mails sent and received in connection with the research), through to highly sensitive (such as financial and medical details of individuals, or details of a major research study of lawbreaking or drug abuse where respondents, who are identifiable, have been assured anonymity). It cannot be stressed too strongly that the degree of sensitivity of the data is irrelevant – all personal data are subject to the Act – but the risk of damage and bad publicity increases with the sensitivity of the data if there is any breach of the Act.
- 3.6.4 The obligations on the HEI and its Cloud Computing supplier are the eight data protection principles, enshrined in Schedule 1 of the Act. HEIs will be familiar with them already. They state that personal data: must be processed fairly and lawfully; that it shall be processed only for specified purposes; that the data should be adequate, relevant and not excessive; that it should be accurate and where necessary, kept up to date; that it should not be kept for any

longer than is necessary; that the rights of data subjects are paramount (see later); that appropriate technical and organisational measures must be taken to ensure there is no unauthorised processing, loss or destruction of personal data (including no unauthorised accessing by third parties to that data); and that personal data may not be moved to a country or countries with inadequate data protection legislation unless full protection of the data is assured.

- 3.6.5 The most important of these principles in respect of Cloud Computing is that data subjects rights must be respected, the data must be protected against unauthorised disclosure, loss, *etc*, and that it must not be transferred to a country with inadequate protection in place. These three are considered further below.

Three key Principles

- 3.6.6 Data subjects, *ie* the individuals who are the subject of the data processing, have the right to inspect the data about them, to know who the data has been disclosed to and where the data has come from, have the right to object to processing of data if they feel it damages them or others, and have the right to sue for any breaches of the Act that has caused them financial damage and/or distress. Thus, the HEI, and its cloud-computing supplier, must be willing and able to provide copies of data to the data subject and to prevent any breach of the Act; they must also keep a record of who has viewed the data (it does not have to be at the level of specific individuals, but broad classes of staff would suffice).
- 3.6.7 The requirement to respond to data subject requests within a tight timeframe is well known in HEIs and there are well-established mechanisms for responding, but the Cloud Computing supplier may not be familiar with them and might be unable or unwilling, for example, to respond to a query from a data subject, or might fail to do so in time. They may also not even recognise a particular request as falling within the Data Protection Act, as the data subject is under no obligation to use the words "Data Protection Act" in any request. This is particularly an issue in respect of US organisations, as there is no Federal Data Protection Act and the companies may not be geared up to responding to requests.
- 3.6.8 The requirement to prevent unauthorised disclosure, loss, *etc* is significant. Whereas it is clearly impossible to guarantee that third parties can never hack into the account (see Information Assurance, below), many Cloud Computing contracts go beyond this and include clauses where the supplier states that it accepts no liability for any loss or destruction of data. Whilst this approach is very understandable from the cloud service supplier's point of view, it leaves the HEI exposed to risk if it accepts this. The Act requires that the data controller – the HEI – imposes obligations on its sub-contractors as onerous as the obligations imposed by the Act on the data controller itself. Therefore, a standard cloud supplier's waiver clause should ring alarm bells for an HEI.
- 3.6.9 Finally, the HEI has potential problems regarding the transfer of data to countries with inadequate data protection laws. The USA is a classic example of a country with inadequate laws, but there are many others. To permit this to happen puts the HEI in potential breach of the Act. Since it is difficult to identify where data is held in a cloud application, the HEI has in effect three choices:
- Insert a clause in the contract that the cloud supplier will abide by all the terms and conditions of the Data Protection Act 1998
 - Insert a series of clauses into the contract specifying the principles that the cloud supplier must follow – these should ideally be worded exactly as in the Data Protection Act. One way the supplier could work with this is to offer a "safe harbour". This is a physical site, perhaps in the USA, where the HEI's data will be kept. The supplier would also need to assure the HEI that the data will not be moved elsewhere and agree that

the space where the safe harbour is will follow the UK Principles (there are standard contractual clauses for this on the Web).

- Insert a clause into the contract confirming that the data will only ever be held in the UK (and/or another member state of the European Economic Area – all have adequate laws). In that way, the data is always subject to the Act (or its EEA equivalent). This is sometimes known as a “local cloud”. The HEI will require a cast iron reassurance that under no circumstances will the data move away from the local cloud.

3.6.10 In summary, current standard Cloud Computing contracts do not offer sufficient “cover” for HEIs regarding their obligations under the Act. HEIs that fail to incorporate the appropriate clauses into their agreements with cloud suppliers could find themselves facing action for a breach of the Act for the failure to impose appropriate obligations on their outsourcing supplier. Suppliers also need to understand the requirements of the Act if they are to sell their services successfully in the UK and elsewhere in Europe. Although many suppliers have signed up for the US/EU Safe Harbour scheme, unless their compliance with the scheme is made contractual, there remains a significant risk for institutions.

Information assurance (IA)

3.6.11 Apart from the legal data protection issues discussed above, funders, institutions, and individual researchers are concerned about the security of their information, although the definitions of security vary. A recent study by the European Network and Information Security Agency (ENISA) provides extensive analysis of the risks and mitigations for Cloud Computing.²²

3.6.12 Many potential users expect cast-iron guarantees that their data cannot be accessed without their authorisation, but it is never possible to give these guarantees. For example, it is reasonable to expect services to protect against common attacks, and to not release user data to the internet. But what about skilled and well-resourced attackers who might be targeting an organisation? New vulnerabilities are constantly discovered in all elements of the internet, and until they are disclosed, they will be exploitable. The real requirement is to make sure that information is protected proportionately to the risk it is under.

3.6.13 The security arrangements put in place by a cloud provider may or may not be adequate for any particular application or dataset. Potential users should apply good risk-management approaches to ensure that their own risk appetite is met. Most cloud providers describe their security approaches publicly, and many have completed some type of external audit. Holding ISO27001/27002 accreditation is regarded as an excellent demonstration of good information assurance policy and practice.

3.6.14 A full treatment of the IA aspects of Cloud Computing is beyond the scope of this document. Some common concerns are described below; it is informative to consider issues against the traditional IA dimensions of Confidentiality, Integrity and Availability.

Confidentiality

3.6.15 Confidentiality is usually the first concern expressed by potential users of cloud services, and may be the only concern that has been considered. There is a perception that there is increased risk in transferring data to an external, usually foreign, service provider, where it will be hosted on a system which is used by many other users simultaneously, and over which the user has no ownership.

²²

<<http://www.enisa.europa.eu/act/rm/files/deliverables/cloud-computing-risk-assessment>> [accessed 21 April 2010]

- 3.6.16 There are undoubtedly some new risks in adopting cloud provision – most obviously, the shift to a hypervised multi-tenant system brings the potential for attacks against the virtualisation layer. If a cloud-based virtual server is compromised, conducting forensics can be very hard – it is not possible to simply turn off the machine and recover the disks for analysis.
- 3.6.17 However, this must be balanced with the concentrations of both risk and expertise within the Cloud Computing providers. These are specialist service delivery and hosting organisations, which have extensive in-house security expertise. Hosting data locally (be it on a personal laptop, departmental server, or university SAN) requires local security expertise that may not be available.
- 3.6.18 Note that hosting virtual servers with an IaaS provider still requires security expertise – although the shared infrastructure may be secure, the security of the virtual server is largely determined by configuration, and that is left to the end user.

Integrity

- 3.6.19 Cloud hosting of data creates new concerns and opportunities for the integrity of data (ensuring that data is not corrupted, either maliciously or accidentally). Cloud providers typically do not conduct backups in the traditional sense, rather they synchronise data between multiple datacentres. Whereas this helps ensure that integrity is maintained, it does not address issues of long-term recovery (historical backups that allow the data as it stood at some point in the past to be recovered), which may be required for some audit activities.
- 3.6.20 For comprehensive assurance of integrity, it would be necessary to host the same datasets on multiple providers (and locally), and conduct regular bit-level comparisons. This degree of reassurance is much greater than most current provision, and is probably unnecessary for the majority of uses.

Availability

- 3.6.21 It is important to define what availability means for any given task. Availability of compute facilities is typically given as an uptime guarantee within a Service Level Agreement (SLA). But the notion of uptime might not be adequate to consider the availability of cloud resources. For example, if an institution's uplink to the internet fails, cloud services will become unavailable to users at the institution. This is outside the control of the cloud service provider, but must be considered. Alternatively, a hosted virtual server may be online (and therefore "up"), but if a hosted database server is down, or the performance of the server is degraded, whilst still remaining up, the service may be compromised. These availability issues require consideration.
- 3.6.22 Although these issues are expressed when considering Cloud Computing, it is evident that they have often not been carefully considered for current provision. Few institutional IT services provide an SLA to their users, and we are not aware of any that match the delivered availability of the major cloud providers.

Contract management

- 3.6.23 It is challenging for any organisation to manage contractual relationships with vendors, particularly when the vendor is very much larger than the organisation itself. Few institutions have the legal and negotiation expertise to contract effectively for mission-critical cloud services. These services are new, and their business models are immature. Standard contracts are typically balanced toward the provider, and for small-scale contracts, it is unlikely that an organisation will be able to negotiate new terms. If an organisation is

considering larger-scale procurement (for example, buying cloud services centrally for use by multiple researchers) there is likely to be more opportunity for variation. We are aware of one major cloud vendor that has altered its contract for SaaS applications to meet the demands of a UK institution – in particular their requirements under the DPA

3.7 Environmental

Green agenda

- 3.7.1 The 'green agenda' is of increasing importance to institutions, and to the sector overall. Incorporation of carbon cutting targets in university HEFCE allocations from 2011-2012²³ (reflecting the Climate Change Act 2008) will force institutions to consider their energy usage directly. It is likely that this will be manifested by energy charges being assigned more directly to the users, which will in turn increase visibility of these charges, and may make the economics of Cloud Computing more favourable.
- 3.7.2 Whether moving to Cloud Computing is actually 'greener' than local provision is difficult to tell. Large data centres are typically located in order to maximise efficiency, for example near to hydropower. Large virtualised data centres can adopt a wide range of tactics to increase efficiency, dynamically powering up or down resources as required. It is likely (although possibly difficult to prove) that utilising large-scale commercial datacentres will reduce the environmental (and particularly carbon) cost of provision.
- 3.7.3 JISC funded a study in parallel with this one, which considered the environmental aspects of Cloud Computing (albeit for administrative computing) in much greater detail than we do here.²⁴

Accommodation

- 3.7.4 Most institutions have significant pressures on the availability of space across their estates. Computer machine rooms have a significant footprint, and have extensive demands for power, cooling and security. The increasing size of compute clusters places increasing demands on machine room space. Moving to cloud provision can reduce the demand for these machine rooms, and potentially free up this space for other uses.

²³ HEFCE 2010/01 – Carbon reduction target and strategy for Higher Education in England.

²⁴ <<http://www.jisc.ac.uk/whatwedo/programmes/greeningict/environmentalreviewcloudcomp.aspx>> [accessed 7 June 2010]

4 Research use cases

Contents	Target audience(s)
Describes the range of use cases for Cloud Computing in research, including case studies of active research, and proposes use cases where cloud approaches may prove particularly beneficial in future. Detailed case studies are included in Annex C.	<ul style="list-style-type: none">– Active researchers– Research managers

4.1 Introduction

- 4.1.1 Any attempt to analyse the computational requirements of entire research areas is fraught with difficulty and generalisation runs the risk of oversimplification. Given this difficulty, and the rapidly changing landscape of cloud offerings, we have chosen to concentrate on providing the tools for researchers and funders to make informed decisions about Cloud Computing for themselves.
- 4.1.2 In doing this, we draw heavily on the information gathered during this project which is presented as examples and case studies. The case studies we present are entirely drawn from science and engineering subjects – we could find no examples of active cloud research in the arts or humanities. This probably reflects the overall balance of experience and effort spent on research computing. Note that the use cases for Cloud Computing (below) are applicable across all disciplines.
- 4.1.3 This section is in no way a comprehensive review of research computing. Indeed, there are an endless number of special requirements researchers may need met. To help researchers relate each example to his, or her, own research problems we have used a broad general classification of the main computational characteristics of each case study.
- 4.1.4 The cases where Cloud Computing may prove beneficial are not closely tied to specific research areas, but rather there may be situations across the range of research computing when it can offer an immediate advantage. The second half of the section draws out and expands upon some areas where Cloud Computing offers significant opportunity and potential for research computing. These use cases are explored more fully in Section 4.4, but in summary they are:
- short timescale requirements;
 - infrequent use and/or no desire to maintain infrastructure;
 - cloudbursting (insufficient local resources);
 - the wrong kind or inappropriate type of local resources;
 - transfer to commercial use;
 - flexibility with system (“root” access);
 - data hosting and backup;
 - cloud-based publications;
 - Ad hoc activities supportive of research.
- 4.1.5 The JISC *Using Cloud for Research: A Technical Review* project presents a range of usage scenarios, which may provide approaches to meeting these use cases.²⁵

²⁵ *Using Cloud for Research: A Technical Review*, (§3.2) Xiaoyu Chen, Gary B. Wills, Lester H. Gilbert, David Bacigalupo, May 2010.

4.2 Computation categorisation grid

4.2.1 In order to perform some assessment of the potential of Cloud Computing beyond the specific examples given in this section we have used a categorisation based on the key characteristics of any computing task. This categorisation is presented in the form of a grid in Table 4-1. The method has the advantage that any researcher with even limited experience should be able to recognise easily where their computing jobs fit in this grid. The options are:

- **CPU time:** is a simple measure of how demanding computationally a task is. It is calculated as: the number of processors × the number of hours taken for the job to run to completion.
- **Degree of parallelisation:** many computational tasks require multiple processors to work together on a problem. If the problem can be split into completely independent tasks that have no need to communicate with each other then this is an **embarrassingly parallel** problem. When the processors performing the task need to communicate with each other then this can add an additional overhead to the overall job completion time. We have defined **fine grained** parallelisation as the case when processors need to communicate very frequently (*eg* as happens in many molecular simulation codes) and **coarse grained** parallelisation when the communication requirement is less frequent (*eg* periodic synchronisation and check-pointing).
- **Data I/O:** is distinct from data passed between processors (which would be covered under parallelisation) and refers to the **volume** and **frequency** with which data is read into the computation and results written out. Since disk I/O is usually a slow operation large volumes of data being frequently written can be a significant factor in the overall performance of a task.
- **Storage:** refers to the *longterm* storage requirement of the work. The issue of storage may seem trivial given the low cost of desktop and external hard-drives but large volumes of data can represent a real problem. Researchers may also be required by their conditions of funding to make data sets available to the wider community. On top of this, the issue of data backup should always be given serious consideration by anybody performing computational research.

4.2.2 Each of the options can take the rating value of **high**, **medium**, or **low** (the option of **very low** has been included for categories relating to data volume, storage and parallelisation. Where a researcher's job characteristics have spanned ratings both are highlighted).

Rating	CPU time / hrs	Degree of parallelisation	Data I/O		Storage
			Volume	Freq	
High	>10,000	Embarrassingly parallel	PB	Constant	PB
Medium	500-10,000	Coarse	TB	Occasional	TB
Low	<500	Fine	GB	Once	GB
Very Low		None	MB		MB

Table 4-1 Categorisation criteria for computational tasks

4.3 Example computational use cases

4.3.1 A range of researchers were contacted and either interviewed or surveyed by questionnaire during the course of this project. Their attitude to using Cloud Computing for research ranged from being enthusiastic proponents of the technology and business model, through

scepticism to “it might be alright for some, but it won’t do for me”. We have presented this experience in the form of ‘use cases’ – coloured **blue** for examples of cloud use, and **red** for examples of general research problems.

4.3.2 The majority of researchers contacted are currently using, or have recently used, Amazon Web Services for their research.²⁶ The range of use cases was broadened to include more of research computing by speaking to researchers who feel that their research requirements are unlikely to be met by Cloud Computing, either for technical or business-model related reasons. The researchers contacted worked in the areas of:

- machine learning in natural language processing;
- flood simulation and decision analysis;
- bioinformatics, including:
 - high-throughput bioinformatics;
 - genetic sequence analysis;
 - protein sequence evolution;
- galaxy formation modelling;
- climate modelling;
- 3D brain imaging;
- website hosting;
- molecular simulation;
- particle physics;
- spectroscopic data processing.

4.3.3 A summary of the characteristics of these researchers’ work (using the categorisation given in 4.2) is shown in Table 4-1. Note that these characteristics are *specific* to the computational tasks of the *researchers we contacted*. This categorisation will therefore not necessarily be the same for all research working in, say, bioinformatics. Rather, the examples given here cover a wide range of computational requirements.

²⁶

The number of AWS examples is due in part to Amazon’s current policy of awarding educational grants, which provided a ready pool of easily contactable researchers. It is also because Amazon is currently the leading supplier of IaaS cloud services. More detail on the different cloud providers can be found in Section 5.

Research area	CPU time	Parallelisation	Data I/O		Storage
Machine learning	Medium – Low	High – Medium	Very Low	Medium	Low – Very Low
Flood simulation	Medium	High	Low	High	Low
High-throughput bioinformatics	Medium	High	Low	Medium	None
Genetic sequence analysis	Medium – Low	Medium	Low	Medium	Medium
Protein sequence analysis	Medium	Medium	Very Low	Medium	Very Low
Galaxy formation modelling	Low	Medium	Very Low	Medium	Very Low
Climate modelling	High	Low	High-Medium	Medium	High
3D brain imaging	Low	High	Low	Medium	Low
Website hosting	None	None	Low	High	Medium
Molecular simulation	High	Low	Low	Medium	Medium
Particle physics	High	High	High	Medium	High
Spectroscopic data processing	Low	High	Low	Medium	Low

Table 4-2 Summary of research use case computational characteristics

4.3.4 Some of the examples from Table 4-2 are presented in more detail below.

Large-scale climate modelling

4.3.5 Climate modelling is a computationally demanding task, producing large volumes of data. In addition to being a technically challenging and evolving research area, the researchers also face the challenge of working under the spotlight of publicity in this increasingly politicised field. In particular calls for openness and transparency are coming both from external sources sceptical of climate change, and from scientists eager to regain public confidence by making data available for scrutiny.

4.3.6 The so-called 'Climategate' furore²⁷ – which involved the Information Commissioner's Office and ultimately elicited a parliamentary response²⁸ – highlights how the forces of 'openness' can create a powerful drive to share raw data and computer code.

²⁷ < <http://en.wikipedia.org/wiki/Climategate> > [accessed 21 April 2010]

²⁸ The report by the Commons Science and Technology Committee into the affair is available at < www.publications.parliament.uk/pa/cm200910/cmselect/cmstech/386/386i.pdf > [accessed 16 April 2010]

Generic use case – climate simulation

Climate modelling typically follows a workflow where a set of initial conditions is provided to a model code, which then generates a set of chaotic intermediate data – running the same experiment twice will result in different intermediate data. These data are then analysed, and statistics produced – these statistics should be identical (within experimental uncertainty) for different runs with the same conditions. These codes develop many TB of intermediate data, which are then processed to several GB of resultant statistics.

The nature of the models is that the world is divided into finite cells, and each cell is analysed independently. As the real world is a continuous system, the cells must interact with their neighbours continually, requiring close coupling of the compute nodes.

Researchers working in this field recognise that their requirements are well met by institutional and national HPC provision, and that cloud services would be both significantly slower and very expensive due to the large data transfer.

CPU time/hrs	Parallelisation	Data I/O		Storage
		Volume	Frequency	
>10,000	Embarrassingly parallel	PB	Constant	PB
500-10,000	Coarse	TB	Occasional	TB
<500	Fine	GB	Once	GB
	None	MB		MB

Machine learning and natural language processing

- 4.3.7 The open availability of much text on the web provides substantial opportunities for researchers attempting to generalise models of natural language and understand the structures underlying knowledge. Specific examples include Wikipedia, datasets explicitly produced for internationally competitive efforts such as the various tasks of the Text Retrieval Conference, and derivative data collections such as Google Web1T, delivered as 24GB of compressed data on 6 DVDs.
- 4.3.8 Typical tasks cover filtering from large collections using information retrieval techniques, and the application of statistical and linguistic models to discover patterns and structures within and across texts, and contrastive analysis across multiple languages. Machine learning techniques can be computationally demanding, with tasks such as the training of artificial neural networks for classification purposes creating demands on both compute and memory.
- 4.3.9 There have been moves towards the provision of cloud-available text processing engines such as "GATE in the Cloud",²⁹ and distributed text processing efforts have a reasonable legacy. Indeed, Hadoop (MapReduce) was designed to be able to reliably handle access to, and processing of, data at such scales, as demonstrated by the petabytes stored and processed by Facebook using Hadoop (now as part of Hive³⁰).

²⁹ GATE in the Cloud: <<http://gatecloud.net/>> [accessed 19 April 2010]

³⁰ Facebook's Hive: <www.facebook.com/note.php?note_id=89508453919> [accessed 19 April 2010]

Cloud use case – Machine learning

Researchers working in the Machine Learning group at the University of Cambridge used **Amazon's Elastic MapReduce** service to allow them to quickly process data sets that would have taken weeks on their local workstations. The computational requirements of their work were relatively modest, however, their code was written in .NET for **serial Windows** machines. This **prevented them using the local Linux cluster**.

They experienced a **steep learning curve** to using the cloud service, which required them to rewrite their code into a suitable language. However, their technical backgrounds enabled them to use the large amount of online material and tutorials to overcome this. Having put the initial effort into learning how to use the cloud they are **positive and enthusiastic** about its potential in their work.

CPU time/hrs	Parallelisation	Data I/O		Storage
		Volume	Frequency	
>10,000	Embarrassingly parallel	PB	Constant	PB
500-10,000	Coarse	TB	Occasional	TB
<500	Fine	GB	Once	GB
	None	MB		MB

Monte Carlo simulation codes

4.3.10 Monte Carlo (MC) based methods are an important and widespread computational tool in fields as varied as:

- finance and economics;
- the physical sciences;
- the life sciences;
- social sciences.

4.3.11 The concept of the 'ensemble' – where the outputs of many simulations are combined in order to make statistical sense of the results – is important in many areas that use MC. This is interesting because, while individual MC simulations can vary in size and may be tightly-coupled parallel processes, the process of generating the ensemble is an embarrassingly parallel task.

Cloud use case – Flood simulation

Researchers at Newcastle University initially used cloud to overcome limitations in the locally available resources. They are now using cloud as a start-up environment in which to develop and demonstrate an approach to bridging the practicality gap between the methods of flood risk simulation and analysis and the resources available to the responsible agencies and their consultants.

Their simulation and analysis packages of flood inundation are computationally demanding and require a significant amount of intermediate storage. However, the problem is embarrassingly parallel; consisting of multi-run computational experiments driven by a probability distribution. The team had been **frustrated by the reliability and performance issues of their Condor grid**, and since their code was provided as Microsoft compatible binaries and relied on SQL Server they **could not use the local Linux cluster**.

The **experience of using the cloud was “largely frictionless”** – although there were some issues with reclaiming cloud services charged to a researcher’s personal credit card!

CPU time/hrs	Parallelisation	Data I/O		Storage
		Volume	Frequency	
>10,000	Embarrassingly parallel	PB	Constant	PB
500-10,000	Coarse	TB	Occasional	TB
<500	Fine	GB	Once	GB
	None	MB		MB

Bioinformatics

- 4.3.12 Bioinformatics melds the principles of information science with molecular biology and is firmly established as a vital tool in many areas including genetic sequencing, protein structure alignment and prediction and drug discovery. It relies heavily on computational searching algorithms and draws on information held in databases. Although bioinformatics can be a computationally demanding area, it is usually thought of as being an inherently HTC rather than HPC process.
- 4.3.13 The amount of information being generated by experiments in molecular biology is growing at a ferocious pace. As the volume of information grows so too does the drive to share datasets. Bioinformatics can make use of ‘standard’ software packages and datasets, and can be amenable to workflow approaches.

Cloud use case – High-throughput bioinformatics

The AptaMEMS-ID project brings together an interdisciplinary team with the aim of developing a handheld device that can identify and distinguish between bacterial strains, such as MRSA. To do this the device identifies surface proteins that are unique to each bacterial strain. Cloud infrastructure is being used to supplement the grid technologies that run the high-throughput bioinformatics studies, which identify the target proteins.

A variety of bioinformatics codes need to be run for each job – which consists of many serial tasks – including BLAST, InterProScan and SignalP. Input files can be several GB in size, making this storage intensive as well as being computationally intensive. EC2 was used to **provide ‘cloudburst’ support to the local Condor grid**, with up to 150 instances running at once.

CPU time/hrs	Parallelisation	Data I/O		Storage*
		Volume	Frequency	
>10,000	Embarrassingly parallel	PB	Constant	PB
500-10,000	Coarse	TB	Occasional	TB
<500	Fine	GB	Once	GB
	None	MB		MB

*Local storage used.

Molecular simulation

- 4.3.14 There are large numbers of researchers working in the field of molecular simulation, which is broadly split into materials modelling and biological simulation. Simulators are large consumers of computing resources at all provision scales – from workstations to HECToR. This is partly because the issue of scale in simulations is in one sense directly related to, and sometimes limited by, the computational power available. Larger machines can run larger (and longer) simulations.
- 4.3.15 Simulations on the molecular scale tend to have the characteristic that they are tightly coupled processes (meaning *fine-grained* parallel processes in our categorisation), utilising MPI or OpenMP to pass information between processors. The traditional platform for this kind of work has been low-latency clusters, but how important it is to use a cluster depends on factors including the size and detail of the simulation and specific code being used.

Generic use case – Molecular simulation

Researchers at the Universities of Warwick and Sheffield took on the challenge of modelling bio-mineralisation processes at the molecular scale. Bio-mineralisation is the process by which organic organisms produce unusual crystals. A good example is the formation of eggshells, which are 95% calcium carbonate with a matrix of proteins that control how the crystals are laid down and prevent the shell from becoming too brittle. Studying this phenomenon by molecular simulation required both the development of sophisticated simulation methods and the use of the HECToR supercomputer.

In order to simulate crystal nucleation the team turned to the technique of metadynamics, which they implemented in the popular simulation code DL_POLY. The simulations of 100,000 atoms showed how the proteins, calcium carbonate particles and water interact to form the eggshell.

The team see this kind of problem as being beyond the reach of Cloud Computing the size and power of HECToR was critical in being able to perform this work. In addition, the team benefitted from being able to optimise the code for HECToR's architecture. The support team at HECToR was able to increase the performance of the code significantly by optimising for HECToR's filesystem.

CPU time/hrs	Parallelisation	Data I/O		Storage
		Volume	Frequency	
>10,000	Embarrassingly parallel	PB	Constant	PB
500-10,000	Coarse	TB	Occasional	TB
<500	Fine	GB	Once	GB
	None	MB		MB

- 4.3.16 Clearly not all simulation is performed at the terascale, indeed the majority makes use of clusters at the institutional level and below. At the lower end of the scale – simulations using a dozen nodes or so – Cloud Computing could offer an alternative resource pool. However, researchers will likely need to judge their requirements carefully.

Overall assessment

- 4.3.17 The overall summary of the responses from the researchers contacted in this projects was, in terms of our categorisation:
- **CPU time:** None of the researchers expressed any concerns over the number of CPU hours consumed in their work on the cloud, either from a technical or financial standpoint. Indeed, several commented that it was reasonably priced. One issue that was raised was that the maximum number of CPUs, or instances, that can be used at any one time on EC2 is capped. However, the cap can be raised with agreement from Amazon.
 - **Degree of parallelisation:** All the researchers we contacted who had real experience of conducting research using Cloud Computing had either *medium* or *high* parallel problems, while the researchers with tightly-coupled problems felt that a low-latency cluster was the most suitable option. This distinction broadly fits with the definition of HTC and HPC respectively (see paragraph 2.1.4). At the level of detail considered in this work this is the only significant technical impediment to using Cloud Computing for research.³¹

³¹

The *Technical Review of Cloud Computing* goes into greater technical detail on the capabilities of Cloud Computing.

- **Data I/O:** None of the respondents reported any issues with disk-write performance. However, the impact of disk I/O is likely to be highly job-specific.
- **Storage:** Most users made use of the free on-instance storage during the runtime of their computations. According to one researcher, the capacity of this on-instance storage compared very favourably with the per-node storage on their local Condor grid. Several groups then stored the final outputs on the much more robust S3 service.³² Cloud-based storage, and associated data transfer, was the largest portion of the cost of Cloud Computing for most of the researchers. The high cost (although whether this is merely a perception or a reality is still in question) of keeping, and downloading, large volumes of data in the cloud is likely to be a barrier to research that produces a lot of data *eg* particle physics. However, this is not a technical limitation but rather a constraint imposed by the business model.

4.3.18 In answer to the question ‘What research can be done with Cloud Computing?’ the shortest answer is perhaps ‘everything that is not tightly coupled’; these problems are better served by traditional HPC facilities.

4.3.19 Technical suitability is, however, only part of the story. A second set of ‘use cases’ emerged from the researchers’ experiences which is less related to the *characteristics* of the research than it is to the circumstances or *scenarios* in which the research must be undertaken. The next section explores these additional use cases.

4.4 Research Use Case Scenarios and possible opportunities

4.4.1 The following Research Use Case Scenarios (RUCS) describe some of the situations in which Cloud Computing may provide significant advantages. There are two basic kinds of advantage: increasing capacity, or increasing capability.³³ These scenarios draw on the use cases within the Cloud Computing White Paper, and are extended by considering the kinds of Cloud resource being used, and by identifying a variety of common features involved. The requirements discussed are described in detail in the Cloud Computing White Paper.

4.4.2 **Security** and **Identity** are common requirements across all use cases, so are not discussed further although the nature of these requirements will likely vary in each case. **Federated Identity** will be a common requirement wherever multiple researchers are concerned.

4.4.3 Scenarios are described in isolation, but these may combine in various ways for specific undertakings. The JISC ‘*Using Cloud for Research: A Technical Review*’ project presents a range of usage scenarios, which may provide approaches to meeting these use cases.³⁴

RUCS.1: Short time scale requirements

4.4.4 **Description:** Specific research needs to be undertaken that has a short-duration requirement for resources. For example, it would take longer to purchase and deploy infrastructure internally, negotiate for and learn to use an internally or externally provided cluster, or wait for a queue on a specific shared cluster, than it would to undertake the research.

³² See subsection 5.4.

³³ In this context, capability is meant as the ability to do something new, or to do something in a new way. In HPC, capability is often meant to mean changes in compute capacity.

³⁴ *Using Cloud for Research: A Technical Review*, (§3.2) Xiaoyu Chen, Gary B. Wills, Lester H. Gilbert, David Bacigalupo, May 2010.

- In this scenario, the research can be expedited more quickly with either an external Cloud service or using an internally virtualised infrastructure (private Cloud) than would be the case with typical infrastructure.

4.4.5 **Service models:** The user will be building and running applications in the Cloud, using Infrastructure or Platform.

4.4.6 **Applicable UCS:** End User to Cloud; Private Cloud

4.4.7 **Requirements:**

- **Location awareness:** Legal or regulatory compliance may restrict where data can be located if using a public Cloud, or which Cloud providers can be used.
- **APIs:** Common APIs should be supported.
- **Benchmarks:** The user may desire to know the performance per unit cost of specific cloud services. This may be helpful in estimating likely costs of any billed services, subject to metering and monitoring.
- **Measured Service:** The user is likely to need to justify costs in order to reclaim expenditure for public Cloud use, or to fit with internal accounting over compute use in a private Cloud.
- **SLAs:** The researcher should be aware of the promises of availability and other provisions in the service level agreement.
- The *ad hoc* nature of the work reduces the likelihood of needing to be concerned with issues such as **Interoperability** or **Portability**.

RUCS.2: Infrequent use and/or no desire to maintain infrastructure

4.4.8 **Description:** Users only require occasional use to servers or cluster systems, potentially wanting to package up what is running on their own machines and run it elsewhere. Alternatively, the overheads of maintaining infrastructure with a relatively low level of utilisation are too high.

- In this scenario, the researcher obtains access to a system configured with whichever software (and versions), scheduling, and so on, as they require. Use of a cluster for MapReduce would be a particular example.
- This is a variant of ***RUCS.1*** since the nature of the work is more likely to be of a generic or systematic nature, for example the infrequent use of a specific service.

4.4.9 **Service models:** The user will be building and running applications in the Cloud, using Infrastructure or Platform.

4.4.10 **Applicable UCS:** End User to Cloud; Enterprise to Cloud; Private Cloud

4.4.11 **Requirements:**

- As ***RUCS.1*** plus the following:
- **Management and Governance:** Enterprises will need to ensure that policies, contracts, and SLAs are negotiated with providers and that the Cloud provision is flexible

enough to ensure that even infrequent needs can be catered for without needing to adopt yet further approaches.

- **Interoperability** or **Portability** amongst Cloud providers, potentially to allow for **Changing Cloud Providers** in order to continue to ensure best value for money.

RUCS.3: Cloudbursting

4.4.12 **Description:** Additional compute capacity is required on demand at specific times, to cope with unpredictable peaks of research computing, than is available within the system which the researcher is currently able to access.

- In this scenario, the researcher is building and running applications with requirements that cannot necessarily be fully satisfied by a particular Public or Private Cloud, using Infrastructure or Platform. These applications will typically be scaling out from a Private Cloud to a Public Cloud, but bursting out across Public Clouds is also a possibility.
- This is a variant of **RUCS.2** in that the nature of the requirement is unpredictable and it should be possible without researcher intervention.

4.4.13 **Deployment models:** The user will be building and running applications in the Cloud, using multiple Infrastructures and/or Platforms.

4.4.14 **Applicable UCS:** Hybrid Cloud

4.4.15 **Requirements:**

- As **RUCS.2**, except:
- **Portability and/or Interoperability:** VM images must either be directly usable across multiple providers, or only incur a small overhead during a necessarily automatic transformation between image formats.
- **Standards:** Standards-conformant APIs are needed to avoid additional costs in creating vendor-specific interfaces and, if possible, to avoid unnecessary data transfer from closely-coupled storage systems or from ephemeral storage.

RUCS.4: Transfer to commercial use

4.4.16 **Description:** A system is required to be accessible, usable and manageable by, or directly transferrable to, a commercial partner.

- In this scenario, the researcher is building and testing applications in the Cloud that, once tested, will also be usable and modifiable by a third party in the Cloud. Neither organisation needs to develop or maintain the infrastructure. Specifying a shared platform becomes a simple task without impinging on the extant infrastructure of either organisation.
- This is a variant of **RUCS.1** in that wider consideration of identity is necessary to provide for additional users. In addition, billing should be transferable away from the researcher.

4.4.17 **Service models:** The user will be building and running applications in the Cloud, using Infrastructure or Platform.

4.4.18 **Applicable UCS:** Enterprise to Cloud to Enterprise

4.4.19 **Requirements:** As *RUCS.1*, except as follows:

- **Portability and/or Interoperability:** VM images may be required by the consuming enterprise for governance reasons, so should either be directly usable across multiple providers, or only incur a small overhead during a necessarily automatic transformation between image formats.
- **Measured Service:** Costs may be billed to the consuming enterprise, in which case metering and monitoring will be essential.

RUCS.5: Flexibility with system ('root') and/or avoiding effects of system upgrades

4.4.20 **Description:** Users want to be able to install and run their own applications without negotiating with system administrators; alternatively, users want to offset the potential up-stack impacts of system upgrades and patches, downtime, and related system factors.

- In this scenario, the user is running legacy server-based applications with specific platform dependencies, or building and testing applications that require full control over everything from the operating system upwards. Such applications need to be impervious to organisational maintenance and upgrade schedules.
- This is distinct from the requirement of a single hosted VM by having requirements for failover or scaling and persisting to storage. A variant of this would be a combination with *RUCS.2* where applications are infrequently tested in multiple configurations with setup and teardown automated.

4.4.21 **Service models:** The user will be building and running applications in the Cloud, using Infrastructure or Platform.

4.4.22 **Applicable UCS:** End User to Cloud; Enterprise to Cloud; Private Cloud

4.4.23 **Requirements:**

- As *RUCS.2*

RUCS.6: Data hosting (and backup)

4.4.24 **Description:** Research organisations wish to make use of the levels of redundancy provided by Cloud storage, and/or retain off-site backups. Alternatively, individual researchers, research groups, or wider research communities wish to have data available “where the computing power is”.

- In this scenario, the researcher is using Cloud Storage.
- Subsequent use may be made of Infrastructure or Platform (*RUCS.1, RUCS.2, RUCS.3* or *RUCS.4*) in relation to the data in Storage.

4.4.25 **Applicable UCS:** End User to Cloud; Enterprise to Cloud

4.4.26 **Service models:** *Cloud Storage*.

4.4.27 **Requirements:**

- **Common APIs for Cloud Storage:** researchers need to be able to access their own data and, where necessary, to enable others to be able to access it. A Content Delivery Network (CDN) that can put data nearer to researchers may be beneficial for pure hosting purposes.
- **Fast Network Connectivity to Cloud Storage:** The frequency with which data are to be transferred, as well as the cost of storage and data transfer, and the ability of the end-to-end network capacity to cope with the demand, will be of key importance to the success.
- **Shipping to Storage Provider:** A provider may need to be able to receive a hard disk that could be copied over to the appropriate Cloud storage.

4.4.28 **Example Scenarios:** Sloan Digital Sky Survey (180GB), Human Genome Data (350GB) and Wikipedia Data (500GB), hosted in Amazon S3.

RUCS.7: Cloud-based research publications

4.4.29 **Description:** To support the repeatability of science, researchers publish papers that contain a reference to VMs and storage that comprise the data, applications, and other essential elements of the published experiment. This enables other researchers to verify fully any claims made about the research, and to undertake further experiments with the same data without each researcher, and supporting technicians, needing to overcome the various barriers to entry to undertaking such research. The VMs and data also form a necessary part of the peer review process.³⁵

- This scenario is a variation on ***RUCS.6*** in which the researcher is using Cloud Storage for both data and VM images.
- Others may subsequently make use of Infrastructure or Platform (***RUCS.1***, ***RUCS.2***, ***RUCS.3*** or ***RUCS.4***) in relation to both the data and the VMs in Storage.

4.4.30 **Service models:** *Cloud Storage*.

4.4.31 **Applicable UCS:** Enterprise to Cloud to Enterprise; Hybrid Cloud

4.4.32 **Requirements:** As ***RUCS.6*** with the exception of:

- **Portability and/or Interoperability:** VM images will be required by the peer reviewers and further researchers, so should either be directly usable across multiple providers, or only incur a small overhead during a necessarily automatic transformation between image formats.

RUCS.8: Ad hoc Activities Supportive of Research

4.4.33 **Description:** Researchers occasionally need to make relatively *ad hoc* use of web-based software in order to undertake efforts that either cannot be supported readily within the organisation due to policy limitations or manpower availability, or to work around specific policies. An obvious example is a limit being placed on the size of files (*eg* images) that can be emailed to project partners, or wanting to host websites outside the confines and templates of an institutional content management system.

³⁵

There is an important proviso here, in that reproducing academic work using the same code will reproduce any errors that are present in the code. For true reproducibility, it is necessary to create a new implementation of the code, using the published methodology/algorithm.

- In this scenario, the researcher will be using SaaS applications, such as Google Mail/Apps that support collaborative activities, or may be using PaaS or IaaS to host websites.

4.4.34 **Deployment models:** The user may be making use of software via all three deployment models.

4.4.35 **Applicable UCS:** End User to Cloud

4.4.36 **Requirements:**

- As *RUCS.1*.

4.4.37 Example Scenarios:

- Google Summer of Code: <http://socghop.appspot.com/>
- Interactive Text Search: <http://quranytopics.appspot.com/> and <http://quranytopics.appspot.com/>;
- The US Environmental Protection Agency's code demonstration system for Watershed Assessment, Tracking & Environmental Results (WATERS): <http://waterssamples.appspot.com/>
- Facebook group for Computational Linguistics Applications 2010: <http://www.facebook.com/group.php?gid=212819479326>

This page is intentionally blank

5 Current cloud offerings

Contents	Target audience(s)
Outlines the services offered by the market leaders. Analyses the contracts and SLAs offered by the three leading providers – Amazon, Google and Microsoft. Further information, including information on other providers is included in Annex E.	<ul style="list-style-type: none">– Active researchers– Research managers– Those making strategic decisions regarding policies or investments– Institutional risk managers, including contracts and compliance officers

5.1 Introduction

- 5.1.1 There is a wide, and increasing, number of providers of “Cloud”, and it is open to debate whether a particular offering comprehensively matches the set of characteristics that NIST, Gartner, or any other organisation might use to distinguish a “pure Cloud” offering from some technology simply rebadged to appeal to a new audience. The entity purchasing the Cloud services needs to decide what the required features are, and try to ignore the overarching labelling.
- 5.1.2 It is apparent, however, that the commoditisation of software, software development and systems can offer alternative ways of doing “traditional” activities. Indeed, for some, IaaS, PaaS and SaaS could be used to replace a significant majority of extant IT systems – entire businesses could run “in the Cloud”.
- 5.1.3 This section provides an overview of market-leading providers of technologies that, in common parlance, are either considered “Cloudy” in terms of what they provide, or relate to providing Cloud for organisations. We leave it to the reader to determine whether such offerings match with expectations of what Cloud *is*. Further details, including information on a much broader range of suppliers, is provided at Annex E.
- 5.1.4 More information about the technologies of Cloud Computing is available in the JISC report on *‘Using Cloud for Research: a Technical Review’*.

5.2 Datacentre as a Service (DaaS) providers

- 5.2.1 Since Cloud Computing is largely reliant on “reselling” data centre capabilities, it is important to consider the role of providers of Data-Centre as a Service, although their full treatment is out of scope for this report.
- 5.2.2 Data centres can make significant advantage of economies of scale to reduce the cost-per-unit service provision and maximize the opportunities for efficient energy use in power and cooling systems. For Cloud Computing users, care needs to be taken in selecting multiple Cloud providers in the hope of constructing redundant systems. It is possible that the physical infrastructures that underlie some of these systems are highly co-located.
- 5.2.3 Information about data centre location and service provider is important to ensure adherence to the Data Protection Act 1998. In particular, where US locations are identifiable, Safe Harbor listings should be sought and examined. Other arrangements are necessary for locations outside the EEA and Argentina, Canada, Guernsey, Isle of Man, Jersey, Switzerland,

Israel, Andorra and the Faroe Islands; Binding Corporate Rule authorisation³⁶ has only been authorised to 6 corporates, none of whom are considered Cloud providers for the purposes of this report.

5.3 Data/Storage as a Service Providers

JungleDisk

5.3.1 **JungleDisk** is a RackSpace subsidiary that offers monthly-charged online storage, with data encrypted prior to upload using AES-256. The storage itself is provided by interfacing to and reselling RackSpace Cloud Files and Amazon S3, although pricing information for each variety of storage presently appears to indicate Amazon S3 fees only (both US and EU).

Caveat emptor

5.3.2 Cloud services are immature, and there are examples of failures in storage providers in particular:

- Sandisk was offering a flash drive, the Cruzer Titanium Plus, with a 6 month free trial of the online backup service BeInSync. BeInSync was based on Amazon S3. Following acquisition of BeInSync by Phoenix Technologies, the BeInSync service was discontinued.
- Oosah promoted 1TB of free storage for media files, but appears no longer to be in existence.
- XDrive offered 5GB free storage, but now only offers pointers to Box.Net and ElephantDrive

5.4 Amazon Web Services: IaaS

5.4.1 Amazon's Web Services (AWS) probably represents the market-leading approach in providing Cloud Computing, with a wide variety of product offerings and a range of further provision, for example ElephantDrive and JungleDisk, being built over AWS.

5.4.2 AWS has a wide range of services. Those most relevant to the scope of this report are:

- **Amazon Elastic Cloud Compute (Amazon EC2):** self-service management of virtual servers which can be built up from preconfigured Amazon Machine Images (AMIs) which comprise a specific operating system and base level of software.
- **Amazon Simple Storage Service (Amazon S3):** persistent data storage accessible in real time via a web services API;
- **Amazon Elastic Block Storage (EBS):** unformatted (virtualised) SAN-like disk of up to 1TB per volume that is independent of virtual servers instances/images;

5.4.3 Other services which may have utility include: Amazon Simple Queue Service (message queuing to enable co-ordination amongst services); Amazon CloudFront (Content Distribution Network); Amazon SimpleDB (a non-relational schema-free attribute-value pair data collection storage service); Amazon Relational Database Service (Amazon RDS); Amazon Elastic MapReduce (a hosted version of the Hadoop implementation of MapReduce).

³⁶

A set of rules that allow an organisation to transfer personal data within their group of companies. The rules are authorised by the Information Commissioner.

- 5.4.4 AWS provides a very close match with the NIST characteristics and Gartner attributes of Cloud Computing, which many other such IaaS providers are not necessarily able to achieve.

5.5 PaaS Providers

Google App Engine (GAE)

- 5.5.1 At the time of writing, the most popular PaaS is probably the Google App Engine.³⁷ This is largely used for developing and hosting websites, which can contain both static and dynamic content, with a programmatic interface provided by either Python or Java using Google's libraries.
- 5.5.2 It is possible to create up to 10 free applications, presented to end users via appspot.com.³⁸ Management of these applications is undertaken by the platform. Programs and data can "roam" around the various Google data centres depending on the level of demand in any country at any time of day, to cope with variations in demand, and to deal with specific demands on capacity. Applications can scale automatically to meet demand.
- 5.5.3 The operating system and database system are selected by Google, and abstracted away from - the so-called "heavy lifting". Certain restrictions are placed on the functionality of applications, such as being unable to open arbitrary network connections.
- 5.5.4 The GAE Software Development Kit (SDK) allows applications to be developed and tested locally prior to deployment. Once deployed, the Google App Engine dashboard allows the application to be monitored, and the dataviewer provides access to data stored and the ability to query over these data using the SQL-like Google Query Language.
- 5.5.5 Only the very largest organisation could undertake such efforts for themselves any cheaper than Google can, given the economies of scale encompassed in the server infrastructure of the latter.

Microsoft Windows Azure

- 5.5.6 Following a free community preview, Microsoft has released the Windows Azure platform. Azure offers compute and storage in Microsoft's data centres, largely intended for hosting .NET applications. Currently, there are three key components to this platform: the Windows Azure operating system, the SQL Server-based Azure database and AppFabric for which SDKs are available to hook in Ruby, PHP and Java programs and use REST and SOAP.
- 5.5.7 The setup required for a development machine, which appears to rule out developers who have not yet upgraded from Windows XP or are constrained from doing this, is relatively substantial. It may be worth using an IaaS provider to obtain a base Windows Server 2008 virtual machine in order to begin work with the platform.
- 5.5.8 In a similar way to Google App Engine, it is possible to run an Azure application on the "local" machine. A development fabric is provided that demonstrates how the Web Roles and Worker Roles of the application will run.

³⁷ <<https://appengine.google.com/>> [accessed 21 April 2010]

³⁸ See <<http://code.google.com/appengine/casestudies.html>> [accessed 21 April 2010] for specific examples of the use of App Engine.

5.6 Derivative Services

RightScale

- 5.6.1 RightScale³⁹ provide a Cloud Management Platform, a single web interface, that can be used to configure and deploy across *multiple* Clouds and across multiple services from Cloud providers, including combined use of Amazon's EC2, S3 and SQS services, Flexiscale and GoGrid.
- 5.6.2 Beyond the free Developer Edition, there are several monthly pricing plans that each include a certain number of "RightScale Compute Units", though it is not made clear what these actually are.

5.7 Private Clouds

- 5.7.1 A number of options exist for an organisation wanting a Private Cloud. Options include Open Source offerings such as Eucalyptus, Nimbus⁴⁰ and OpenNebula and paid-for private provision from the likes of VMWare,⁴¹ Enomaly,⁴² Platform Computing,⁴³ Intalio⁴⁴ and Surgient.⁴⁵
- 5.7.2 A variation on this theme is the hosting of a Private Cloud by a public provider as exemplified by Amazon VPC. Amazon VPC⁴⁶ provides a mechanism for using part of a public Cloud as if it were a private ("hosted") Cloud. Amazon provides secure connectivity to a set of AWS resources, via an IPsec VPN that is charged for by the hour, that are isolated from other AWS resources, which remain in the public Cloud.
- 5.7.3 In this Section, we provide brief overviews of Eucalyptus and OpenNebula.

Eucalyptus

- 5.7.4 Eucalyptus⁴⁷ is a reverse-engineered implementation of Amazon EC2 and S3 that emerged from the University of California Santa Barbara and the VGrADS project,⁴⁸ and is now provided by Eucalyptus Systems in two forms: an Open Source project from which people can build their own Private Clouds, and a corporate venture that builds and maintains Private Clouds.
- 5.7.5 Eucalyptus is mostly written in Java. It implements a set of web services that are interface-compatible with certain services of Amazon AWS. Communication amongst the components of Eucalyptus is also undertaken via web services. Eucalyptus supports both REST and SOAP, and each service has an XML definition of the services on offer.

39 <<http://www.rightscale.com>> [accessed 24 April 2010].

40 <<http://www.nimbusproject.org/>> [accessed 21 April 2010]

41 <<http://www.vmware.com/>> [accessed 21 April 2010]

42 <<http://www.enomaly.com/>> [accessed 21 April 2010]

43 <<http://www.platform.com/>> [accessed 21 April 2010]

44 <<http://www.intalio.com/>> [accessed 21 April 2010]

45 <<http://www.surgient.com/>> [accessed 21 April 2010]

46 <<http://aws.amazon.com/vpc/>> [accessed 21 April 2010]

47 <<http://open.eucalyptus.com/>> [accessed 21 April 2010]

48 <<http://vgrads.rice.edu>> [accessed 19 April 2010]

5.7.6 Example users:

- NASA’s Nebula is a \$2m pilot project. Eucalyptus is a major element of this provision. While there is some technical information available, aside from the fact that it is located within a 40ft data centre container (shown lower right in <http://nebula.nasa.gov/static/nebula/nebula-container.jpg>), there is little information available regarding the scale of the infrastructure.⁴⁹
- The NGS has been undertaking initial efforts in Cloud Computing, utilising Eucalyptus.⁵⁰
- Eucalyptus is a key part of the £0.5m investment in the St Andrews Cloud Computing initiative to support PhD research. The Cloud infrastructure provides for up to 64 virtual machines.⁵¹
- An implementation of Eucalyptus providing for up to 64 virtual machines is also available at the University of Surrey. This private Cloud is supporting PhD research, and being used alongside Amazon’s EC2, and complementarily to Google App Engine, as part of an MSc module in Cloud Computing.⁵²

OpenNebula

5.7.7 OpenNebula⁵³ is a virtualisation system, provided as Open Source, which emerged from the Distributed Systems Architecture Research Group at Complutense, University of Madrid.

5.7.8 OpenNebula can use KVM, Xen and VMWare to run virtual machines, and can be used in combination with the Haizea lease management system to provide for reservations and scheduling based on specified service level requirements.

5.7.9 Whereas Eucalyptus attempts to implement the full set of EC2 API calls, OpenNebula provides only a subset of these. In particular, storage makes use of more “traditional” mechanisms, and does not appear to be provided in a service-oriented manner, as it is in Eucalyptus. The subset provided enables EC2 to be used to launch and manage server instances in order to provide for Hybrid capability.

5.7.10 Example users:

- Amongst other sources, development of OpenNebula is receiving EU support through the Resources and Services Virtualisation without Barriers (RESERVOIR) project⁵⁴.
- CERN have reported using OpenNebula as a key part of a wider implementation to manage about 400 servers, with a peak load of 7500 virtual machines⁵⁵.

49 <<http://nebula.nasa.gov>> [accessed 23 April 2010]

50 <<http://www.ngs.ac.uk>> [accessed 23 April 2010]

51 <<http://www.cs.st-andrews.ac.uk>> [accessed 23 April 2010]

52 <<http://www.eps.surrey.ac.uk/cloud>> [accessed 23 April 2010]

53 <<http://www.opennebula.org>> [accessed 21 April 2010]

54 <<http://www.reservoir-fp7.eu/>> [accessed 21 April 2010]

55 <<http://lists.opennebula.org/pipermail/users-opennebula.org/2010-April/001886.html>> [accessed 21 April 2010]

5.8 Contracts and SLAs

5.8.1 Each provider has its own approach to contracting with clients, and these contracts and SLAs are immature. It is important for institutions to consider carefully the relationship that contracts with cloud providers will establish, and to consider the implications of an SLA. Bear in mind that most institutional and local provision at present has no SLA. SLAs set out a minimum quality of service, and in most cases, suppliers significantly exceed the performance of their SLA.

Consideration of current contracts

5.8.2 We examined the following contracts and SLAs, to consider their current suitability for UK academic institutions:

- Amazon contract⁵⁶
 - Amazon S3 SLA⁵⁷
 - Amazon EC2 SLA⁵⁸
- Google App Engine Terms of Service⁵⁹
- Windows Azure Terms of Use⁶⁰
 - Azure SLAs for compute and storage⁶¹

5.8.3 This subsection sets out our opinion on these contractual documents. It should not be used as a basis for forming a contract or otherwise. It does not constitute legal advice, and does not establish a solicitor-client relationship.

Recommendation 1: any organisation considering adopting any cloud services for mission-critical applications, or for processing personal or otherwise sensitive information, should obtain specialist legal advice regarding their contracts and SLAs.

Overview

5.8.4 These are typical agreements of their type. As such, they are one sided in favour of the service supplier and against the client. Of the agreements, the Microsoft one is, in our opinion, the most user-favourable. The style is very much lawyer-speak, and there is a risk that HEIs, or non-specialist staff within them, will sign without realising all the implications. Briefly, the services agree to supply a service to the clients for a fee, but only accept limited liability if things go badly wrong and they do not accept basic legal obligations on their operations.

5.8.5 One particular issue to note is that these contracts are established under US, rather than any UK law, and require conflicts to be settled in US courts. Our analysis below highlights some areas where the contracts may be illegal under UK law, but may be permissible under their own jurisdiction.

56 <http://aws.amazon.com/agreement/>

57 <http://aws.amazon.com/s3-sla/>

58 <http://aws.amazon.com/ec2-sla/>

59 <http://code.google.com/appengine/terms.html>

60 <http://www.microsoft.com/licensing/onlineuserights/english>

61 <http://www.microsoft.com/windowsazure/sla/>

Amazon Web Services

- 5.8.6 The Amazon EC2 service level agreement offers a 10% credit on the bill if uptime percentage falls below 99.95% in a year. This refund does not increase as downtime increases. It might be more equitable to offer 10% of the fee or the % of downtime, whichever is the greater, but it doesn't. Furthermore, the credit is against future bills. But what if the client has terminated the contract - say because it was so disgusted with the 50% downtime? The client would then get no money back at all - it is just credit notes against future payments. Even after all that, Amazon excludes all liability for downtime that results from any actions or inactions by the client or a third party - that is pretty much an all-embracing waiver, as Amazon could use this as an excuse for any downtime that might occur. And there's another exclusion for any downtime caused by failures of individual instances - so instances (and possibly the whole service that a client depends on) may become unavailable, but this does not count as downtime. Finally, Amazon states it will look at each case in turn and at its own sole discretion decide whether to award a refund or not. This SLA Agreement could be described as extremely unfair (but not necessarily illegal - *caveat emptor*).
- 5.8.7 Interestingly, the Amazon S3 service level agreement offers credit if downtime is worse than 99.9% rather than 99.95% as noted in the first one, and offers up to 25% credit. This is somewhat more generous in terms of credit and eligibility, but the same caveats apply to how Amazon can get out of its commitments as for the EC2 service. So, in practice, it is not that much better.
- 5.8.8 The general Amazon contract has several features which would require detailed consideration before accepting:
- Amazon can change a free service to a priced service at no notice, or to change its fees at any time;
 - the client agrees to be bound to revised terms and conditions even though the client is not informed that there are such terms and conditions changed - it is up to the client to keep checking the web site to identify any changes;
 - Amazon can suspend the service at any time for any or no reason, giving just 60 days notice; if the service is suspended, the client is still liable for ongoing charges even though the client cannot use the service;
 - Amazon claims it is not responsible for any unauthorised access to client data, even though the Data Protection Act says that a service supplier *is* liable when the data is personal data (in this case the DPA would probably over-ride the contractual term, but it could be messy legally if such an incident occurred);
 - Amazon states that the client is not permitted to reverse engineer or decompile its software, but Clause 50B of the 1988 Copyright Designs and Patents Act says it IS legal for a lawful user to do this under some circumstances and the Act makes it clear that any contractual clause attempting to restrict this right is null and void. At the very least it should be saying "to the extent permitted by law, you may not.....";
 - the client is required to assign copyright in any "Text Materials" (not properly defined) to Amazon - this would need clarification in a negotiation with Amazon;
 - there is a catchall clause allowing Amazon downtime for maintenance or service modifications or for other reasons, and the client is NOT entitled to any refund and Amazon is not obliged to warn clients of any downtime scheduled;
 - Amazon reserves the right to pass over any information given to it by a client to a Government agency on request (this presumably is primarily US Government for anti-terrorist reasons, but could be any other Government for any reason the Government might make);

- Amazon requires the client to warrant that nothing in its content is illegal, a warranty that many institutions would find difficult to absolutely guarantee - "to the best of your knowledge" should be inserted;
- there is nothing regarding Force Majeure for the clients - what if due to a power cut their payment system crashes and so they don't pay Amazon on time? Amazon itself has all sorts of Force Majeure get outs, but clients are not offered this.

Windows Azure

- 5.8.9 The SLAs for Windows Azure services are in many ways similar to those provided by Amazon. For both computing and storage, they offer credits for service failures, but these do not cover events beyond MS' control, where the cause is the client's hardware or software, or if caused by any action by the client or a third party. This, of course, could always be used as a reason not to offer refunds. Credits cannot exceed the fees payable, and can only be used to offset future bills. Problems caused by upgrades and the like carried out by MS are excluded from consideration for the discount. Certain levels of problem trigger either 10%, or 25% off the next monthly bill, and so, for example, if the service was down for 50% of a month, the client could still only get a maximum of 25% off their next bill.
- 5.8.10 In comparison to Amazon, the Windows Azure contract is an improvement in a number of ways. For example:
- it promises to e-mail clients of changes to the contract and assures them that if they don't like the changed terms, they can continue to run under the old terms for at least 12 months;
 - it promises not to track, view, censor or edit client data (Amazon is silent on this point);
 - it promises to use reasonable endeavours to ensure there is no unauthorised loss or disclosure of a client's data.
 - The terms and conditions for suspension or termination are reasonable.
- 5.8.11 There are more minor issues regarding this contract:
- Microsoft will access or disclose data if required to do so by a government.
 - There is nothing about Force Majeure in the main contract.

Google App Engine

- 5.8.12 The Google App Engine contract has several issues:
- It says anyone over 13 can enter into such a contract, but in English law to enter into a legal contract the person must be at least 18.
 - Like Amazon, credit is for future purchase and cannot be a cash refund.
 - Google also must respond to any Government request for disclosure of data.
 - Unlike the others, Google reserves the right to edit, filter, remove, *etc* content placed by a Client.
 - Google accepts no responsibility for deletion of, or failure to store, content supplied to it by a Client.
 - The client takes full responsibility for the security and back up of data given to Google.
 - Google has the right to use the client's name in publicity, including logos, domain names, *etc*

- If the service itself is changed, Clients can still use the old version but it will no longer be maintained or upgraded.
- Google reserves the right to terminate a Client's use of the service for a variety of reasons, including if the service provided to the client is not profitable to Google! The client then has 90 days to transfer the content elsewhere.
- Google can change the terms and conditions at any time without having to communicate them to the Client - it's up to the Client to keep checking on the Web whether the terms have changed.
- Force Majeure only applies to Google and not the client.

General comments

5.8.13 All three contracts fail to cover:

- 1) The obligations of the service supplier to abide by the Data Protection Act if personal data is being supplied to the Cloud Computing service.
- 2) Who is responsible if an illegality occurs (*eg* copyright infringement, defamation, pornography, race hate materials or incitement to terrorism) when the data has been amended by the Cloud Computing service provider and whose laws would be applied?
- 3) How would a cloud service provider respond to a "three strikes and you are out" law (*eg* the Digital Economy Act)?
- 4) Cloud suppliers should build privacy and security enhancing features into the product and service design, *eg*, automatic encryption of all data held (not offered by these suppliers); personal data is not held, or is only held for short periods, or is always anonymised; Privacy Impact Assessments should be carried out by independent experts and the contract should permit Clients to demand them if they are not on offer (and they are not in the three cases examined).
- 5) The contracts should cover issues such as:⁶²
 - i) Who can see my information?
 - ii) Who is responsible for data corruption or loss?
 - iii) How easy would it be to migrate to a competitor cloud service supplier?
 - iv) Where will our data be held?⁶³
 - v) What individuals are responsible for managing our data?
 - vi) What security technology is being employed?

Alternatively, they could provide a contact point where such questions can be answered. Indeed, all the contracts should commit the supplier to providing a help desk, but none of them do.

⁶² Some of these are addressed within the service descriptions from the various providers, but it would reduce the risk to institutions if they were handled contractually.

⁶³ All of these providers are signed up for the US/EU Safe Harbour registration scheme, which enable international transfers that are compatible with EU (and UK) data protection legislation. However, this safe harbour registration is not included within the contract.

This page is intentionally blank

6 Analysis and conclusions

Contents	Target audience(s)
Synthesises the range of information collected during this study, and presents the overall view of current and possible future uses of Cloud Computing in research;	– All readers

6.1 Introduction

- 6.1.1 This report is scoped to focus on infrastructure clouds – the provision of virtual servers and storage. Whereas Cloud Computing creates some new issues, the vast majority of applications which could utilise cloud infrastructure must be considered in a similar manner to current local provision. For example, asking how to manage backup in the cloud is a similar question to asking how to manage backup in an institutional datacentre. The answer varies depending on the specifics of the services being provided, the organisational strategy, the availability of funding, the skills and interests of the individuals involved in specifying, supporting and managing the service, and many other factors.
- 6.1.2 The biggest risk, and perhaps the biggest opportunity for any plans utilising cloud services is the current immaturity and hence the rapid growth and development of the market. The biggest players are multibillion dollar companies with established track records, so potential users of these services must consider that although there is a possibility that their chosen vendor may cease trading or cease provision of a service without warning, this is exceedingly unlikely. The futures of the smaller organisations are less secure.
- 6.1.3 In this section, we highlight some of the cloud-specific issues that have been raised elsewhere in this report, and suggest approaches to consider or address these issues where possible. Nonetheless, we cannot provide answers to how to use clouds – only ideas to consider.

6.2 Current situation

- 6.2.1 Cloud Computing is currently at an early stage of adoption within higher education, and even more so within research computing. Those who are using cloud facilities for research are typical for early adopters of any new technology – they have technical backgrounds, they enjoy investigating new technologies, and they are self-reliant problem solvers, and they are mostly addressing “hard science” problems.
- 6.2.2 At present, the vast majority of research computing use of Cloud Computing has been funded by Amazon.⁶⁴
- 6.2.3 There is interest amongst the academic community about the opportunities for Cloud Computing, but a general sense amongst the communities that we have engaged that it is “too early”. We understand this to mean that the cloud technologies and business models are not yet sufficiently mature to give research computing users the confidence to commit the significant time and effort required to move much of the existing large-scale research computing toward the cloud. Typical research grants run for 1-3 years, which is a very long time in the development of a new technology. Committing to use cloud resources for research presents the very real possibility that the resources that are available at the end of

⁶⁴ Note that this report is focused on infrastructure (compute and storage) provision. Clearly, SaaS applications such as webmail are widespread.

the project, are different (and perhaps less suitable) from those which were available at the beginning, with the consequent difficulties in ensuring continuity of access to the systems upon which the research will by then depend.

- 6.2.4 There is some resistance to change amongst academics, but this resistance is for the most part well considered and argued – it is not naïve obstructionism. There are significant barriers to the large-scale uptake of Cloud Computing for research, most significantly:
- Having compute resources which are free at the point of use is beneficial for training and innovation;
 - The costs of institutional research computing infrastructure provision are hidden from the users of that provision, being subsidised by the host institution. This leads to increased costs directly to the researcher in a cloud environment;
 - There are significant disruption costs to the researcher in moving existing research computing tasks to the cloud;
 - There are likely to be performance penalties in moving to virtualised delivery of many research computing codes (see below).
- 6.2.5 The research councils and other funding agencies could potentially drive the uptake of cloud by explicitly asking bidders to consider whether any requests for compute resources could be met by shared or on-demand provision. There is, however, a risk that bid assessors lack the expertise to review statements made by bidders in response.

Applicability of cloud technologies to research problems

- 6.2.6 Cloud technologies could theoretically meet almost any research requirement, but as with any technology, the fit will depend on the exact requirements of the task. It is worth noting that the use of Cloud Computing in research computing is novel enough that significant findings are considered research in their own right, and usually appear within the academic rather than the professional literature.
- 6.2.7 The majority of research computing representatives considered IaaS to be the service model that was most applicable to their needs – on-demand rental of compute or storage, with the flexibility to control these blank systems however desired. This is probably due to the conceptual leap from locally owned machines to on-demand virtual machines being relatively straightforward.
- 6.2.8 Other service delivery models may prove more applicable to mature codes that require less modification during the course of the research project. This may be considered to be the distinction between research into the codes or the methodological approach, and research that depends on the codes as tools (for example, using BLAST⁶⁵ to search protein sequence databases).
- 6.2.9 Licensing software for use on cloud services does not present any particular challenges for research codes. The range of licence terms applied by different vendors preclude general advice, but it is clear that as Cloud Computing becomes more established within research, licences will be adapted to refer specifically to cloud services. It is also possible that software vendors will adapt by offering their software as a service.
- 6.2.10 Two major areas have emerged as being currently **unsuitable** for migration to a cloud platform. These are situations in which an extremely large amount of data needs to be processed/stored, and for large-scale fine-grained parallel jobs. However, this is an

⁶⁵ <<http://blast.ncbi.nlm.nih.gov/>> [accessed 24 April 2010].

assessment made on *current* cloud provision only, and may well be subject to rapid change as vendors bring new services and offerings into production.

- **Large-scale data handling and processing** is currently likely to prove unaffordable due to charging models used by commercial cloud service vendors.
- **Fine-grained/tightly-coupled parallel processing** for example using OpenMP or MPI is currently likely to suffer unacceptable performance overheads from virtualisation into a cloud.

6.2.11 That said, the management of research data is a significant current problem. Cloud provision provides opportunities for storage and management of research data, albeit with a new set of problems regarding sustainability, longevity, and cost.

Recommendation 2: JISC should investigate the issues surrounding the management and preservation of research data in cloud systems, and produce guidance aimed at researchers. This should support risk assessment and management, and should not design or develop technical solutions.

6.2.12 The most significant current opportunities for Cloud Computing in research are in the support of new capabilities, for example a user with a code that runs in Windows, who occasionally needs a more powerful machine for specific tasks. This user could, with relatively little effort, rent a powerful virtual windows machine when required, rather than invest capital in a more powerful machine.

6.2.13 Although we made efforts to engage with research computing users within the Arts, Social Sciences and Humanities, the number of individuals active in this area is very limited, and the majority of these are using workstation computing rather than large-scale provision. These users could potentially benefit from the flexibility of IaaS clouds to give them more of what they already have, without the significant challenges of migrating to Unix or Linux to operate on clusters or HPC machines.

Current cloud services

6.2.14 There is a wide range of cloud services available, although the only one that has made any noticeable impression on research computing is AWS, and this is primarily due to their extensive programme of funding research using their services. However, Google and Microsoft have significant cloud infrastructures and are aggressively marketing to academia, for administrative computing at least. Both Google and Microsoft have interest in the use of their systems for research, so it seems likely that they will grow in relevance to research computing.

6.2.15 AWS is distinct from Google and MS in that it offers IaaS, whereas the other two offer PaaS. IaaS is a more flexible base upon which to build research stacks, allowing the user to determine the nature of the virtual machines that they use, and to exert full control over them. This flexibility appeals to the current, early adopters of Cloud Computing for research, rather than the bulk of research computing users, who tend to value stability. As usage increases, this may change.

Data management and preservation

- 6.2.16 Cloud Computing generates new considerations for the backup and archiving of information. This can be for information that is stored or created on the cloud or for information that is created or exists elsewhere that could be backed up or archived on a cloud service.
- 6.2.17 Infrastructure clouds do not provide backup in the traditional manner. The integrity of data in the cloud is provided by real-time replication between multiple datacentres, rather than offline or nearline backup. Requirements for preservation of historical data must be met by the systems running on the cloud infrastructure for example by replicating data locally or to another cloud provider.
- 6.2.18 A significant issue regarding the reliability of cloud services for the preservation of data is the widely held but frequently incorrect feeling amongst users that local data, whether on a removable hard disk in a researcher's office, on a local server, or elsewhere in the institution, is more secure than that held in a remote datacentre. The ability to see and touch the device that the data is stored on is remarkably reassuring!
- 6.2.19 From a security (confidentiality) aspect, many potential users expect cast-iron guarantees that their data cannot be accessed without their authorisation, but it is never possible to give these guarantees. For example, it is reasonable to expect services to protect against common attacks, and to not release user data to the internet. But what about skilled and well-resourced attackers who might be targeting an organisation? New vulnerabilities are constantly discovered in all elements of the internet, and until they are disclosed, they will be exploitable.
- 6.2.20 The real requirement, for confidentiality, integrity and availability of data is to make sure that information is protected proportionately to the risk it is under, and the consequences should the data be disclosed, damaged or made unavailable.

Leveraging the sector scale to negotiate service levels

- 6.2.21 The current services and business models are immature, and there are issues regarding data protection, information assurance, contracts, procurement and stability. It is clear that individual institutions (or researchers) will have limited opportunity to negotiate with the major cloud vendors. It may be possible to deliver broad benefits by establishing a central mechanism for negotiation (*eg* a framework contract), and possibly for procurement.

Recommendation 3: JISC should investigate mechanisms for national engagement, negotiation and procurement of cloud services, primarily with AWS, Google and MS, but allowing for the engagement of smaller, niche providers.

6.3 Outlook

- 6.3.1 It is important to distinguish between the use of cloud technologies to provide new capacity for research computing, or new capability.⁶⁶ Given the present barriers, it is unlikely that there will be spontaneous uptake of Cloud Computing to provide new capacity, except in some very limited cases (very short-term requirements or the need for flexibility). There is much greater opportunity in meeting the requirements that are not currently well served by research computing provision (some examples of which are set out in sub-section 4.4).

⁶⁶ In this context, capability is meant as the ability to do something new, or to do something in a new way. In HPC, capability is often meant to mean changes in compute capacity.

- 6.3.2 Cloud Computing is already established as an important mechanism for the delivery of services within industry, and increasingly within university administration. In time, it is likely to become an important part of the research computing infrastructure. This is likely to be through the provision of new capabilities initially, rather than by displacing local or national facilities, which serve their user communities well already.
- 6.3.3 It is clear that institutions should consider their approach to Cloud Computing – is it acceptable for individual researchers to use these services without institutional consideration of the risks?

Staff skills

- 6.3.4 The skills required to administer computing facilities and those required to develop applications that run on these facilities are different at present, and will remain so as Cloud Computing is adopted.
- 6.3.5 The current infrastructure cloud services essentially allow local machines to be replaced by virtual, on-demand machines. This may require some specific changes in the skills of research computing services (as any new technology does), but this skills change is relatively minor, taking a period of days or possibly weeks to become familiar with these technologies. It does not represent a fundamental change in the skillset.
- 6.3.6 An interesting possibility is that, over time, new generations of researchers will consider Cloud Computing as a natural alternative to local clusters. This may lead to the development of algorithms that are better suited to the nature of cloud provision, and which will in turn lead to greater uptake of Cloud Computing.
- 6.3.7 The notion of systems architecture is different in cloud services, with systems being created to meet *ad hoc* requirements; the flexibility and dynamic nature of clouds goes against the traditional stable systems architecture paradigm. There is currently an effort to increase the software engineering skills of researchers, in part to increase the suitability of research codes to operate on traditional HPC. This movement is likely to continue, and will provide researchers with the skills to utilise cloud services (as well as other resources) effectively – in time.

Opportunities for support

- 6.3.8 There may be opportunities for Cloud Computing to allow researchers to experiment with codes, which whilst well established, are not available to them on local resources. Whilst it would be possible to install and configure codes to experiment with, there is a potential benefit in the provision of trusted machine images for common scientific codes. The NGS seems to be a suitable organisation to host the development and dissemination of such images.

Recommendation 4: The NGS, and its funders, should consider whether there is a role for that organisation in supporting the development of virtual machine images for common research codes, to allow users to deploy them easily within commercial and private clouds. This may include liaising with or funding the developers or maintainers of the codes.

A UK academic cloud?

- 6.3.9 There has been some discussion of the potential for creating a UK academic cloud, either centrally provided, or by federating institutional resources in a similar manner to the NGS.

Such an arrangement is argued to be beneficial by reducing costs (through the lack of a profit margin), and increasing flexibility and “ownership”.

- 6.3.10 For research computing, these arguments are not persuasive. Cloud Computing is made efficient and effective by its very large scale. This scale permits the flexible and elastic nature of the services that are provided. A UK academic cloud would not have the scale to realise the key benefits of Cloud Computing, yet would still accrue most of the disbenefits.
- 6.3.11 A UK academic cloud would have a real risk of becoming a system that has the appearance of a cloud, some of the functionality of a cloud, the same interfaces, but not the capacity or capability of a true cloud. It is likely to remain the preserve of those who are interested by the technology, rather than those who want to use the technology to meet their research requirements. The experience of NGS has shown that even the provision of services that are free at the point of use is not sufficient to overcome disbenefits of shifting to new ways of working. Small-scale clouds for development and testing could be provided locally, but the economics compared to buying capacity from commercial suppliers will need to be carefully considered.

Recommendation 5: unless backed by clear evidence of demand, and a robust and revenue-neutral business case, JISC should not support the development of a production UK academic research cloud.

- 6.3.12 This is not intended to advise against continued investigation of cloud technologies for research, but the focus should be on the development of research codes to make best use of *all* available resources, including the cloud, rather than on trying to fit cloud into existing research problems.

7 Recommendations

7.1 This section lists the recommendations made throughout the document.

- **Recommendation 1:** any organisation considering adopting any cloud services for mission-critical applications, or for processing personal or otherwise sensitive information, should obtain specialist legal advice regarding their contracts and SLAs.
- **Recommendation 2:** JISC should investigate the issues surrounding the management and preservation of research data in cloud systems, and produce guidance aimed at researchers. This should support risk assessment and management, and should not design or develop technical solutions.
- **Recommendation 3:** JISC should investigate mechanisms for national engagement, negotiation and procurement of cloud services, primarily with AWS, Google and MS, but allowing for the engagement of smaller, niche providers.
- **Recommendation 4:** The NGS, and its funders, should consider whether there is a role for that organisation in supporting the development of virtual machine images for common research codes, to allow users to deploy them easily within commercial and private clouds. This may include liaising with or funding the developers or maintainers of the codes.
- **Recommendation 5:** unless backed by clear evidence of demand, and a robust and revenue-neutral business case, JISC should not support the development of a production UK academic research cloud.

This page is intentionally blank

A Interviews conducted

A.1 The following individuals and organisations have provided their time and expertise in the support of this project, and the authors would like to extend their greatest appreciation and thanks for this assistance.

Name	Organisation
Andy McGregor	JISC
Andy Richards	NGS
Chris Rawlings	Rothamsted Research
Daniel Perry	JANET(UK)
David Carr	Wellcome Trust
David McAllister	BBSRC
David Quigley	University of Warwick
Frances Collingborn	NERC
Frederique van Till	JISC
Hamish Harvey	University of Newcastle
Iain Gavin	Amazon Web Services
Ian Stewart	University of Bristol
Jane Nicholson	EPSRC
Jeremy Neathy	ESRC
Jon Blower	University of Reading
Jurgen van Gael	University of Cambridge
Keith Flanagan	University of Newcastle
Mark Baker	University of Reading
Mark Rodger	University of Warwick
Mark Thorley	NERC
Matt Ismail	University of Warwick
Matthew Cook	University of Loughborough
Michael Daw	University of Manchester
Ned Garnett	NERC
Neil Chue Hong	OMII-UK
Neil Geddes	STFC

Paul Valdes	University of Bristol
Peter Tinson	UCISA
Phil Richards	University of Loughborough
Rudolf Römer	University of Warwick
Sebastian Bratieres	University of Cambridge
Simon McIntosh-Smith	University of Bristol
Steve Gough	University of Reading
Syma Khalid	University of Southampton