

Cloud computing for research

# annexes to final report

CC421D008-1.1

14 June 2010

Cover + 40 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@curtis-cartwright.co.uk](mailto:postmaster@curtis-cartwright.co.uk)  
web: <http://www.curtis-cartwright.co.uk>

Registered in England: number 3707458

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



## Document history

Version	Date	Description of Revision
0.6	28 April 2010	New document separating annexes from main document
1.0	7 June 2010	Release version
1.1	14 June 2010	Post-release corrections

This page is intentionally blank

## List of contents

<b>Document history</b>	<b>iii</b>
<b>List of contents</b>	<b>1</b>
<b>A Interviews Conducted</b>	<b>3</b>
<b>B Cloud computing taxonomy</b>	<b>3</b>
B.1 Characteristics	3
B.2 Cloud Computing Use Cases – White Paper	7
<b>C Case studies</b>	<b>11</b>
C.1 Case Study – Machine Learning	11
C.2 Case Study – Decision analysis for flood risk in coastal cities	13
C.3 Case Study – High-throughput bioinformatics	15
<b>D Survey Responses</b>	<b>19</b>
D.1 Introduction	19
D.2 How much experience did you have with Cloud technologies before the grant?	20
D.3 Why did you decide that Cloud could be useful?	20
D.4 What were the biggest barriers to getting started with the Amazon Cloud (AWS), and how long did you think it took you to get up and running?	21
D.5 Which parts of AWS have you been using in your work - <i>eg</i> EC2, S3, Elastic MapReduce, Cloudwatch	21
D.6 How would you Describe your use of Cloud? (Data set, duration of computation, parallelism <i>etc</i> )	21
D.7 Is there anything you would have like to have done with AWS but could not?	22
D.8 Were there any privacy or other regulatory/legal issues in your work?	22
D.9 Were there any other surprises. Pleasant or otherwise?	22
D.10 Will you continue to pay to use a public cloud at the end of the Amazon funding?	23
<b>E Further details on current cloud providers</b>	<b>25</b>
E.1 Introduction	25
E.2 Storage as a Service	25
E.3 Infrastructure as a Service	25
E.4 Software as a Service (SaaS)	30
E.5 Derivative Services	31
E.6 Platform as a Service	31
<b>F Cloud Computing Research</b>	<b>33</b>
<b>G Cloud-based Datasets</b>	<b>39</b>

This page is intentionally blank

## A Interviews Conducted

A.1 Annex A is contained within the main document.

## B Cloud computing taxonomy

### B.1 Characteristics

B.1.1 The five NIST characteristics are comparable and variously coincident to Gartner's Five Attributes of Cloud Computing.<sup>1</sup> We provide a brief synthesis of characteristics that emerge from the two below.

*NIST: On-demand self-service; Gartner: Service-based*

- **automatic provision without human interaction.**
- **service orientation** is a particular focus for the Gartner definition, with considerations of abstraction to a service, and the importance of service level descriptors.

*NIST: Broad network access; Gartner: Uses Internet Technologies*

- **protocols and standard mechanisms for interacting with the service.**

*NIST: Resource pooling; Gartner: Shared*

- **services may be sharing underlying computational resources** in a way that isolates one consumer from another within an underlying system, for example in a "multi-tenant model" (NIST). In such a system, it becomes possible that competitor organisations may be hosting services on, for example, the same physical server without ever being aware, able to detect, or able to interfere with the presence of each other. Additionally, the services may be migrated elsewhere and the physical server assigned to yet another organisation without any of these organisations being aware of this. These resources may exist across geographical boundaries, and the consumer may or may not have control over where a service runs (NIST).

*NIST: Rapid elasticity; Gartner: Scalable and Elastic*

- **services should be able to scale resource provisioning, automatically and quickly, to meet changing demand**
- **upper limits on capacity should not be the concern of the consumer**

---

1

*NIST: Measured Service; Gartner: Metered by Use*

- **monitoring should be available, and reportable and may be related to service billing**
- **resource use levels should be manageable by both provider and consumer**

B.1.2 Other definitions tend to repeat certain of these characteristics and may add others. For example, Reese<sup>2</sup> defines Cloud Computing as: “a service accessible via a web browser or web services API that requires zero capital expenditure and for which you may pay for what you use as you use it”. This definition identifies the distinction between the up-front investment, and need for balance sheet reporting, of capital expenditure in contrast to operational expenditure. This difference often emerges when discussing the economics of, and advantages of, Cloud Computing.

### ***Service models***

B.1.3 The three service models according to NIST are Software as a Service (SaaS), Platform as a Service (PaaS) and Infrastructure as a Service (IaaS):

- **Software as a Service (SaaS):** Consumers are able to access software that runs online on the systems of the software provider. Consumers may be able to customize the software, or embed it in other systems via specific software development kits (SDKs), but cannot delve into the software to alter its functionality or gain access to the underlying systems to use them for other purposes.
- **Platform as a Service (PaaS):** Consumers have the ability to write programs to specific software development kits (SDKs) which will be able to run on the systems of the platform provider. The consumer can write whatever software they wish within the limitations imposed by the platform provider. The consumer cannot access the underlying systems.
- **Infrastructure as a Service (IaaS):** Consumers have administrative access to systems which appears to allow access to underlying systems and which can allow for all kinds of development without restriction to specific SDKs. Typically such servers would make use of virtualisation (V12N), so consumers only have the impression of being able to run on the underlying system. There may be limitations on the kind of virtualised hardware available on such systems, in terms of network hardware, and important considerations to be made over how to provide for persistent storage.

### ***Software as a Service (SaaS)***

- SaaS is considered the most mature aspect of Cloud Computing, and provides for software that is usable over the internet. SaaS offerings tend to be those that scale readily and easily to large number of (potentially paying) customers, with relatively well-known requirements; email and customer relationship management systems are typical SaaS examples.
- Common software moved from being a boxed product, distributed on media, to becoming a hosted download; this shifts the burden of acquisition, but the burden of installation and maintenance remains on the organisation. SaaS applications, and users

---

<sup>2</sup> Reese, G. (2009) “Cloud Application Architectures: Building Applications and Infrastructure in the Cloud: Transactional Systems for EC2 and Beyond”. O’Reilly Media, Inc. ISBN(13): 978-0596156367, 204 pages.

of SaaS applications, are supported by the organisations who should know best how to offer such support. Issues relating to software “bloat”, whereby downloads become increasingly large, hard to set up, and potentially demanding an upgrade of the hardware, are problems for the provider, not the consumer.

- SaaS applications can be readily updated to resolve specific problems, for example to remove vulnerabilities; this does not rely on specific maintenance schedules within organisations which may leave a certain proportion of systems with vulnerabilities. Costs, where applicable, on a “per seat” basis; these should encompass all the costs of maintenance, upgrades and so on, and account for the level of use of the software.
- **Examples:** Google Mail, and other web-based email systems, and Google Apps are common examples of SaaS.

### ***Platform as a Service (PaaS)***

- PaaS provides a specific environment in which programs coded specifically to that environment can be run.
- PaaS typically involves an infrastructure hosted on the internet, a software development kit (SDK), access to provider-specific data storage, mechanisms for interacting with the platform, and monitoring and billing services.
- The limitations on programming language and constraints imposed by provider platforms may be unacceptable to those with legacy applications; re-engineering an entire application in order to deal with either of these may be a significant hindrance to uptake.
- **Examples:** Google App Engine and Microsoft Windows Azure are common examples of PaaS.

### ***Infrastructure as a Service (IaaS)***

- IaaS provides the possibility for entire computer systems, from the operating system up, to be created and run. The IaaS provider owns the system hardware and networking.
- The consumer is responsible for everything from the guest operating system, or multiple guest operating systems, upwards; licenses, where necessary, are the responsibility of the consumer.
- Typically, such provision implies the use of virtualised infrastructure, where there is an “illusion” of having full control of hardware. However, it should not be possible to tell whether a virtual instance is running alone on a physical machine, or sharing a physical machine with others (multi-tenant). There may be performance degradation in a virtual machine in contrast to a physical machine due to overheads imposed by the hypervisor; however, a good provisioning system should be able to meet specifiable levels of performance.
- It should be possible to have servers provisioned and destroyed on-demand – “self service” – without requiring intervention from the provider.
- IaaS is closest to a Rental model of computing, but should be at a much finer grain in terms of pay as you go (PAYG) that is exemplified by monthly or annual rental agreements.

- **Examples:** Examples of Public IaaS providers include Amazon, Rackspace and GoGrid. Institutions can also become internal providers of Clouds, with private IaaS can be provided using software such as Eucalyptus.

B.1.4 The three models abstract away from the underlying **systems**, and should allow the provider to scale their own systems efficiently.

B.1.5 The notion of **multi-tenant** systems is relevant to Cloud Computing. Such a system allows several consumers to co-exist on the same physical system without ever being aware of, or being able to interfere with, the systems of others.

B.1.6 A wide variety of other “as a Service” labels are in use, including Hardware as a Service (HaaS), Data as a Service (DaaS), Business as a Service (BaaS), and Everything as a Service<sup>3</sup> (EaaS, XaaS, \*aaS)

### **Deployment models**

B.1.7 The four deployment models encompass:

- **Private Clouds:** Refers to the virtualisation of an “internal” data centre using software such as Eucalyptus or Open Nimbus, where only consumers internal to the organisation are able to use.
- **Public Clouds:** Cloud systems available to the general public, offered by providers such as Amazon, RackSpace, GoGrid, Google (App Engine).
- **Community Clouds:** These may be considered as Federated Private Clouds where some organisations share (parts of) “private” Cloud systems with each other, but not with the general public
- **Hybrid Clouds:** Refers to the capability of making combined use of private, public and potentially community clouds depending on the requirements of the application.



*Figure A-1: The deployment models allow for a reasonable degree of interpretation. An organisation may have multiple Private Clouds, one of which can also be used by a Community; a Community Cloud may make use of multiple Private Clouds, and one which enables use of a Hybrid Cloud; Hybrid Clouds could variously combine Public, Private and Community at different times.*

<sup>3</sup> <[http://en.wikipedia.org/wiki/Everything\\_as\\_a\\_service](http://en.wikipedia.org/wiki/Everything_as_a_service)> [accessed 14 June 2010]

## B.2 Cloud Computing Use Cases – White Paper

- B.2.1 The Cloud Computing Use Cases White Paper, version 2.0 presents 7 Cloud Use Case Scenarios (UCS) that generalise how users and enterprises interact with Cloud systems.<sup>4</sup> These help to generalise the Research Use Case Scenarios (RUCS) featured in the main document.
- B.2.2 The White Paper contains one research-relevant application, an “Astronomic Data Processing” customer scenario. The scenario itself describes the European Space Agency’s use of a Cloud, with substantial estimated savings in using a Cloud over an in-house equivalent. The scenario is firstly characterised as “Enterprise to Cloud to End User” (p32), and subsequently as “End User to Cloud” (p37).

### End User to Cloud

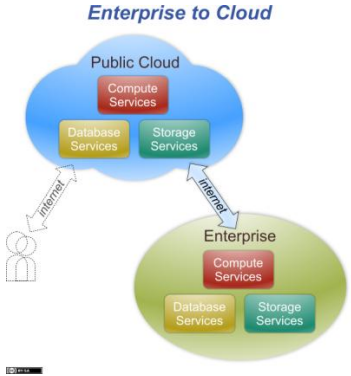
<p>B.2.3 The End User to Cloud UCS is described as “Applications running on the cloud and accessed by end users”.</p> <p>B.2.4 A Customer Scenario relating to this Use Case is not provided in the White Paper.</p> <p>B.2.5 <b>Research Use:</b> SaaS, eg Facebook, LinkedIn; Typical Research Uses of Platform and Infrastructure</p> <p>B.2.6 <b>Examples:</b> Facebook group for Computational Linguistics Applications 2010: <a href="http://www.facebook.com/group.php?gid=212819479326">http://www.facebook.com/group.php?gid=212819479326</a></p>	<p>From: Cloud Computing Use Cases group files.</p>
--	---

### Enterprise to Cloud to End User

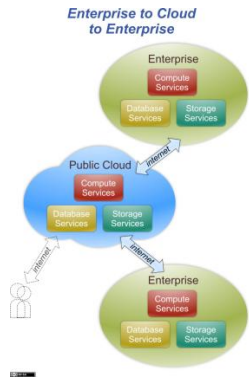
<p>B.2.7 The Enterprise to Cloud to End User UCS is described as “Applications running in the public cloud and accessed by employees and customers”</p> <p>B.2.8 Two Customer Scenarios relate to this Use Case in the White Paper: (i) Logistics and Project Management.</p>	<p>From: Cloud Computing Use Cases group files.</p>
---	---

<sup>4</sup> <<http://groups.google.com/group/cloud-computing-use-cases/files>> [accessed 14 June 2010]

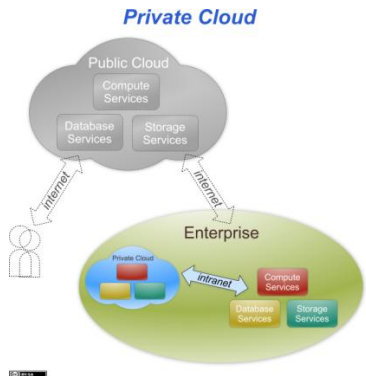
### Enterprise to Cloud

<p>B.2.9 The Enterprise to Cloud UCS is described as “Cloud applications integrated with internal IT capabilities”</p> <p>B.2.10 “This use case involves an enterprise using cloud services for its internal processes”.</p> <p>B.2.11 The Customer Scenario relating to this Use Case in the White Paper concerns Payroll Processing.</p> <p>B.2.12 <b>Research Use:</b> Cloud-hosted (outsourced) organisational Email / Docs / CRM <i>etc</i></p>	 <p>The diagram, titled "Enterprise to Cloud", shows a "Public Cloud" at the top containing "Compute Services", "Database Services", and "Storage Services". Below it is an "Enterprise" containing "Compute Services", "Database Services", and "Storage Services". Bidirectional arrows labeled "Internet" connect the Public Cloud and the Enterprise. A person icon on the left is connected to the Enterprise via an "Intranet" arrow.</p> <p>From: Cloud Computing Use Cases group files.</p>
--	---

### Enterprise to Cloud to Enterprise

<p>B.2.13 The Enterprise to Cloud to Enterprise UCS is described as “Cloud applications running in the public cloud and operating with partner applications (supply chain)”</p> <p>B.2.14 <i>A Customer Scenario relating to this Use Case is not provided in the White Paper.</i></p>	 <p>The diagram, titled "Enterprise to Cloud to Enterprise", shows a central "Public Cloud" containing "Compute Services", "Database Services", and "Storage Services". It is connected via "Internet" to two "Enterprise" nodes, each containing "Compute Services", "Database Services", and "Storage Services". A person icon on the left is connected to the central Public Cloud via an "Intranet" arrow.</p> <p>From: Cloud Computing Use Cases group files.</p>
--	--

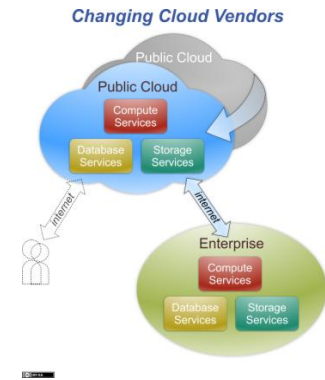
### Private Cloud

<p>B.2.15 The Private Cloud UCS is described as “A cloud hosted by an organisation inside that organisation’s firewall”</p> <p>B.2.16 The Customer Scenario relating to this Use Case in the White Paper concerns Central Government.</p> <p>B.2.17 <b>Research Use:</b> System Benchmarking; Monitoring and SLAs (Surrey)</p> <p>B.2.18 <b>Example:</b> StaCC, OeRC, Surrey FEPS</p>	 <p>The diagram, titled "Private Cloud", shows a "Public Cloud" at the top containing "Compute Services", "Database Services", and "Storage Services". Below it is an "Enterprise" containing a "Private Cloud" (with "Compute Services", "Database Services", and "Storage Services") and another "Enterprise" (with "Compute Services", "Database Services", and "Storage Services"). Bidirectional arrows labeled "Internet" connect the Public Cloud and the Enterprise. A person icon on the left is connected to the Enterprise via an "Intranet" arrow.</p> <p>From: Cloud Computing Use Cases group files.</p>
---	--

**Changing Cloud Vendors**

B.2.19 The Changing Cloud Vendors UCS is described as “An organisation decides to switch cloud providers or work with additional providers”

B.2.20 A Customer Scenario relating to this Use Case is not provided in the White Paper.

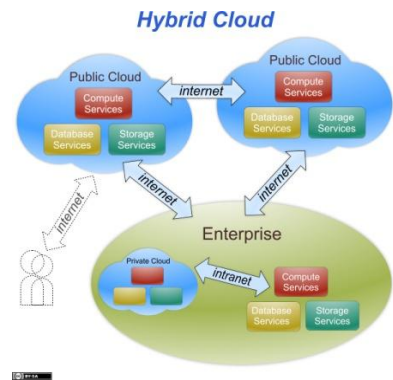


From: Cloud Computing Use Cases group files.

**Hybrid Cloud**

B.2.21 The Changing Cloud Vendors UCS is described as “Multiple clouds work together, coordinated by a cloud broker that federates data, applications, user identity, security and other details”

B.2.22 The Customer Scenario relating to this Use Case in the White Paper concerns Local Government.



From: Cloud Computing Use Cases group files.

This page is intentionally blank

## C Case studies

### C.1 Case Study – Machine Learning

#### *Summary*

- C.1.1 Researchers in Machine Learning used Amazon's Elastic MapReduce service to allow them to quickly process data sets that would have taken weeks on their local workstations. They experienced a steep learning curve to using the 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.

#### *Background*

- C.1.2 Jurgen van Gael and Sebastien Bratieres, Department of Engineering, Cambridge University
- C.1.3 Jurgen works in the Machine Learning group of the Learning Sciences division of the School of Engineering. He is a 3<sup>rd</sup> year PhD student with a background in Computer Science. He is comfortable with several programming languages and is highly technically competent.

#### *Description of research*

- C.1.4 Jurgen's research is in Machine Learning and he is specifically interested in natural language processing. He has written code in .NET that uses a Map-Reduce approach to analysing a data set consisting of single sentences in which each word of the sentence is tagged; each tag denotes the part of speech to which the word belongs. The code takes a data set in which the tags have been generated at random and iteratively improves the tags by applying a computationally expensive algorithm. The jobs also require several GB of RAM per processor.

#### *How and why did you become interested in the Cloud?*

- C.1.5 Jurgen was limited in what he could accomplish on his workstation. The demanding nature of the calculations meant that jobs could take on the order of a week for even relatively small input data-sets of ~10 – 100MB.
- C.1.6 The Learning Sciences division also has its own Beowulf-type linux cluster – the main users are researchers in Speech Analysis who run HTK. The cluster is supported by two systems administrators, who also support the other computational needs in the Learning Sciences division.
- C.1.7 In order to take advantage of the departmental cluster for his work and speed up his jobs Jurgen would have to rewrite his code into a linux compatible language *and* implement parallel programming (which he was not familiar). The code would have to handle the bookkeeping associated with splitting-up the data and allocating processors. Jurgen had neither the technical knowledge nor time to invest in this process.
- C.1.8 Using the Distributed Matlab Toolkit on the cluster was a possibility but the systems administrator informed Jurgen that the cost would be prohibitive.

- C.1.9 There is a distributed version of a MapReduce type operation compatible with .NET available called DryadLinq, but this would require Jurgen to have his own Windows HPC cluster. IBM provided a quote for making such a cluster available remotely, but the cost was unacceptable.
- C.1.10 Jurgen wanted to process larger data sets more quickly than could be done on his workstation. His proposal to Amazon was based on some simple estimates based on scaling his desktop calculations. He was awarded \$5000 for the use of Amazon Elastic MapReduce.
- C.1.11 Reason for choosing Amazon included: the reputation of Amazon, the quality of service expected from a commercial offering. The extensive documentation and tutorials available from Amazon and the range of educational materials on using Amazon readily available on the web were praised.

### ***Details of using the cloud***

- C.1.12 Before they could use the cloud Jurgen and his colleague Sebastien spent several weeks looking at the documentation and learnt Hadoop by using Cloudera's tutorials and emulator that allowed them to run a version locally.
- C.1.13 Jurgen and Sebastien rewrote their .NET code (which is not supported by Amazon) into Java. This took several weeks of effort. They note that debugging the code until it ran smoothly was a serious challenge, although the Amazon engineers were unexpectedly helpful. On one occasion even providing unsolicited advice on why the code had crashed.
- C.1.14 Details of using the cloud:
- Code and input data set deposited in an S3 storage 'bucket'
  - Job submitted: at this time the type of desired instance to be used is specified, as is the location of the code, input data. On job completion the output data is written back to the bucket.
  - Job submission and monitoring is handled through the AWS interface.
- C.1.15 A typical 'job' will consist of:
- 100Mb input file
  - 10,000 iterations of the algorithm; each iteration is a single MapReduce job and takes ~1minute when run on 12-24 processors (translated into instances).
  - Each iteration produces ~100Mb of intermediate data, this is kept on the processing node and discarded.
  - After each iteration some statistics are gathered, which inform and improve the next iteration.
  - Final output is ~100Mb and is written back to S3.
- C.1.16 The processes of adapting code, learning how use Hadoop and the cloud took several months. After 6 months experience Jurgen and Sebastien still say they are 'getting the feel' of how their jobs run on the cloud.

### ***Legal issues***

- C.1.17 Jurgen and Sebastien did not have any data protection requirements on their research.

- C.1.18 They were satisfied that Amazon would provide a high quality of service and uptime. This judgement was based on the reputation of Amazon as a commercial company rather than an examination of the SLA.

### **Economics**

- C.1.19 Jurgen and Sebastien are happy with Amazon's billing system. They did however have one unpleasant surprise early on when their intermediate data was being written to S3 instead of remaining on-node.

### **General comments on using the cloud**

- C.1.20 A testing sandbox would have been nice. Currently all testing is live and billed.
- C.1.21 Jurgen now thinks that rewriting his code into Python rather than Java would have been better.
- C.1.22 Generally, Jurgen is very positive about using the Amazon Elastic MapReduce service. The current preliminary work has led to a paper, and he is very keen to expand to data-sets of ~1Tb which would not have been possible on his desktop.
- C.1.23 Jurgen noted that the systems administrators in his department were interested in the possibilities of the cloud. Changes to the way the institution pays for electricity mean that the individual departments could be incurring more of the costs. The systems administrator estimated that up to 1/6<sup>th</sup> of the departmental energy bill was generated by the local cluster.

### **Classification**

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

## **C.2 Case Study – Decision analysis for flood risk in coastal cities<sup>5</sup>**

### **Summary**

- C.2.1 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

<sup>5</sup> *Uncertainty analysis using Amazon Web Services*, Hamish Harvey and Jim Hall, 9<sup>th</sup> International Conference on Hydroinformatics, HIC 2010, Tianjin, China, submitted.

risk simulation and analysis and the resources available to the responsible agencies and their consultants.

### ***Background***

- C.2.2 Hamish Harvey, Research Associate, School of Civil Engineering and Geosciences, Newcastle University.
- C.2.3 Hamish is a civil engineer by background, but has a strong interest in computing. He began following the development of Cloud Computing out of general interest, but quickly realised the potential it presented to his research problems.

### ***Description of research***

- C.2.4 Hamish runs simulation and analysis packages of flood inundation. The problem is computationally expensive and storage intensive. The computation is large but embarrassingly parallel. It consists of multi-run computational experiments of simulation models of flood inundation. These are driven by samples from a probability distribution or density function. The codes are Microsoft compatible binaries and rely on SQL Server.
- C.2.5 A significant volume of intermediate storage is required, especially as it is often desirable to be able to trace the computation back from final results to the individual simulation runs. These intermediate results therefore need to be kept so they can aid that exploration.

### ***Why use the Cloud?***

- C.2.6 Hamish had for a long time been frustrated by both the limitations of the workstation-bound scale of his research, and by what he perceived to be the high barrier to getting started on clusters. His simulation and analysis codes need to be run on Windows machines, so in any case the local Linux cluster was unsuitable. He did have access to a Condor grid that included Windows machines; however, these were not configured with the required software. A working pool of machines was eventually configured, but reliability and performance issues remained a serious problem.
- C.2.7 Hamish first used AWS following the introduction of Windows instances to rapidly regenerate results for a consultancy assignment after discovering a bug in the analysis code. He was able to deploy the analysis to the cloud and re-generate the results in a week, thereby saving many weeks of waiting. This experience convinced him that the technology was both stable and usable, and that it addressed the problems of computationally expensive and storage intensive research.
- C.2.8 Part of this work involves developing and providing tools to assist users in setting up the analysis and in interpreting the results. The team is working on web-based tools, which fits well with using a cloud backend for the computation.

### ***Details of using the cloud***

- C.2.9 Jobs used EC2 instances for the processing work, S3 for job input and output data and the Simple Queue Service (SQS) for distributing jobs to worker virtual machines.
- C.2.10 A heterogeneous virtual infrastructure was used, in which one or two Linux instances generated jobs and performed post-processing, while up to nineteen virtual Windows

instances (all of the medium size, high CPU type) processed jobs. Worker instances were based on an Amazon-supplied custom image with Windows and SQL Server Express pre-installed, and modified so that upon booting the instance would immediately begin pulling jobs from the queue.

### **Legal issues**

- C.2.11 There are no legal or data protection issues associated with this work.
- C.2.12 Hamish has no concerns about the SLA at this stage in his work since it could “only be compared with the SLA I would give myself!”. However, if the work to develop a widely-used decision analysis tool is successful then he feels the SLA will be more important.

### **Economics**

- C.2.13 When Hamish first started using AWS he did so by providing his own credit card details for his account. This led to some explaining to the departmental finance officer when he tried to reclaim the costs. Since Amazon will only accept credit card payments Hamish feels this could be a potential stumbling block for wider uptake of Cloud Computing.
- C.2.14 The cost of data storage and transfer was the largest part of the cost of using AWS. Six weeks work, in which 6000 machine hours were consumed, came to a cost of USD1500.

### **General comments on using the cloud**

- C.2.15 Overall Hamish described the experience of using AWS was described as “largely frictionless”.

### **Classification**

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

## **C.3 Case Study – High-throughput bioinformatics**

### **Summary**

- C.3.1 The AptaMEMS-ID project<sup>6</sup> 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

<sup>6</sup> <<http://gow.epsrc.ac.uk/ViewGrant.aspx?GrantRef=EP/G061394/1>> [accessed 8 April 2010]

strain. Cloud infrastructure is being used to supplement the grid technologies that run the high-throughput bioinformatics studies, which identify the target proteins.

### ***Background***

- C.3.2 Keith Flanagan, Research Assistant, School of Computing Science, Newcastle University.
- C.3.3 Keith has a background in computer science and is a highly competent and technically skilled user of computers. He has at least a year of experience with AWS stemming from his PhD research.

### ***Description of research***

- C.3.4 A variety of bioinformatics codes need to be run for each job, including BLAST, InterProScan and SignalP. Input files can be several GB in size, so the problem is storage intensive as well as being computationally expensive. However, the problem is embarrassingly parallel since each 'job' consists of many serial tasks.

### ***Why use the Cloud?***

- C.3.5 Keith first used Cloud Computing in his PhD research, in which he wrote middleware controller scripts to allow EC2 instances to run bioinformatics analysis on data transferred directly to the compute node via BitTorrent. He is now applying the same methodology in the AptaMEMS-ID project.
- C.3.6 The bioinformatics group has access to a local Condor grid. However, there were some limitations as to its use. Although the total size of the Condor pool is ~10,000 machines, the overwhelming majority are Windows machines, and the bioinformatics codes need Linux-based machines. Of these, not all have the proper software installed, which left only about 80-or-so machines which were suitable for Keith's purposes. In addition, the regular day-to-day users of the machines were complaining about the machines being slow whilst computations were running.

### ***Details of using the cloud***

- C.3.7 EC2 is used to provide computing power in a 'cloudburst' manner. EC2 instances are requested as needed from the same control script that also submits jobs to the local Condor grid.
- C.3.8 The EC2 instance receives the job command and retrieves the data to be analysed from local servers using BitTorrent. Each 'job' runs on a medium instance for upto 2 hours, and each workpackage can consist of many thousands of jobs. The largest number of instances Keith has had running at any one time on a job is around 150. It was necessary to apply to Amazon to increase the standard cap of 20 instances.
- C.3.9 If necessary instances are controlled either by the Amazon Web-Interface, or the command line.

### ***Legal issues***

- C.3.10 There were no legal or data protection issues with this work.

- C.3.11 Keith attached no real importance to the SLA provided by Amazon. This was, in part, due to the fact that his code had been designed to cope with machine failures and interruptions, which can be a common occurrence on Condor grids.

### **Economics**

- C.3.12 The cost is “cheap, but not insignificant”. It should be noted that all longterm storage requirements for this work are handled by local resources, thereby negating the monthly cost of storage on S3, for example.
- C.3.13 Keith is happy that the technology used by Amazon is reliable and stable, and he would consider applying to use cloud-based resources in future funding proposals to the Research Councils.

### **General comments on using the cloud**

- C.3.14 Keith briefly looked at other cloud providers, and particularly Google App Engine. However, he thought that Amazon (which uses an IaaS model) offered a greater level of flexibility and control of the virtual infrastructure, which fitted with both how his code is implemented and how he preferred to work.
- C.3.15 The maximum allowed size for AMIs hosted by Amazon is 10GB. This proved insufficient to allow a custom AMI to be created with all the required software and packages fully-installed. Keith had to work around this by creating a custom AMI that called a local server to install the relevant codes.
- C.3.16 Vendor lock-in is a concern for Keith, but he is satisfied that the IaaS model offers sufficient flexibility to avoid this.

### **Classification**

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

- C.3.17 Storage needs met by local servers.

This page is intentionally blank

## D Survey Responses

### D.1 Introduction

D.1.1 Fifty researchers from around the world were emailed a short survey regarding their use of Cloud Computing. Their names were gathered from the publically available Amazon site listing details of educational grants. A total of eleven useful responses were received – nine respondents gave direct answers to the survey and two respondents provided papers published based on their work on AWS. The list of respondents, their affiliation and a short project summary is given in Table C-1.

Name	Affiliation	Project
Andres Monroy-Hernandez	MIT, MIT Media Lab	Scaling Scratch, a new visual programming environment for children that makes it easy to create interactive stories, animations, games, music, and art—and share them on the web
Andrew Benson	California Institute of Technology, Theoretical AstroPhysics Including Relativity (TAPIR)	Exploring models for contents of galaxy formations
Christopher N. Hill	MIT, Department of Earth, Atmospheric, and Planetary Sciences	Mathematical models for analyzing complex Earth ocean systems
David Quigley	UCSF Cancer Research Institute	Analysis to identify genes that function in normal lung biology as well as in lung tumor pathology
Geoffrey Charles Fox	Indiana University, Pervasive Technology Institute	Proof of concepts linking FutureGrid users to AWS
Greg Aldering	University of California Lawrence Berkeley National Lab, The Nearby Supernova Factory	Understanding the origin of supernovae, how they explode, and how to better calibrate them as distance indicators
Jiang Dawei	National University of Singapore, Computer Science	MapReduceDB: A unified data processing system
Mark Pearrow	MIT, McGovern Institute	Genetic and computational analysis, electrophysiological recordings, and non-invasive brain imaging
Michael C. Schatz	University of Maryland, Center for Bioinformatics and Computational Biology	Assembly of large genomes using Cloud Computing
Ralph Mietzner	University of Stuttgart, Institute of Architecture and Application Systems	Automated deployment and management of composite applications distributed across cloud computing environments.
Renquan Cheng	School of Information Systems, Singapore Management University	New techniques in malware analysis

Saptarshi Guha	Purdue, Department of Statistics	Install scripts and cost performance analysis for RHIPE and Hadoop via EC2
Taylor Sittler	University of California San Francisco, Viral Diagnostics and Discovery Center	Deep sequencing for influenza virus detection and discovery
Till Quack	ETH Zurich, Computer Vision Lab	Large scale annotation of photo collections
Zachary Ives	University of Pennsylvania, Computer and Information Science Department	Orchestra, collaborative data sharing system on the cloud

Table C-1 survey respondents

D.1.2 This section describes the results of this survey. The dataset is too small to draw meaningful statistics, but the answers provide useful snapshots of the benefits and challenges of Cloud Computing.

## D.2 How much experience did you have with Cloud technologies before the grant?

D.2.1 Respondents had between zero and 2 years experience in using AWS.

*"Not much experience other than reading about the different offerings from Amazon and attending one of their local events in Boston."*

*"[our collaboration] had absolutely zero previous experience with Cloud computing prior to the grant. However, we worked with members of the Advanced Computing for Science Department here at [...] who have experience with Cloud technologies and especially with Amazon Web Services. This was critical to getting up and running as they were very familiar with the AWS interface and knew what to expect in terms of system behaviour, performance, and even provided us software tools to configure our virtual 'cluster.'"*

## D.3 Why did you decide that Cloud could be useful?

D.3.1 Several interesting responses were received in answer to this question. These have been used in the synthesis of the cloud use scenarios in the main document. Evaluation of cloud technologies for the wider scientific audience was also mentioned as a specific reason for seeking the AWS grant.

*"The initial appeal of cloud has been the on-demand access and the ability to package an environment as a complete system."*

*"I was undertaking a project which needed significant amounts of computing power for a period of a few months. The type of computing needed was large number of serial calculations so I wasn't concerned about using a cluster for parallel codes. I'd heard about Amazon's EC2 and thought this might be a good use for it..."*

*"Our motivation was to explore an alternative to the "scientific computing centre" model we had been using – a large Linux cluster at a laboratory or university...Over the years, the management of that cluster has upgraded operating systems and hardware in the Linux cluster which had consequences for our software pipeline – the net result is that our pipeline developers (astrophysicists, not computer scientists) have to stop doing science and handle the resulting code ripples. This is something we felt shouldn't be a*

*priority for our personnel, and the obvious alternative (running our own mid-scale cluster) was definitely cost-prohibitive in terms of both initial acquisition and maintenance...It should be stressed, however, that we went into our experiment with EC2 fully expecting to learn that even though the specific costs of both the "scientific computing centre" model and running our own private cluster could be avoided, EC2 would probably pose \*new\* complexity costs."*

#### **D.4 What were the biggest barriers to getting started with the Amazon Cloud (AWS), and how long did you think it took you to get up and running?**

- D.4.1 Getting started was a barrier to some respondents, although other respondents specifically mention how easy getting up and running was. Other barriers mentioned include: cost, benchmarking, reduced performance compared to dedicated resources, and creation of AMIs.

*"The biggest barrier was probably figuring out how to set up access credentials and to configure my computer to connect to EC2. But, it wasn't actually that difficult. I found reasonably good instructions online and shortly afterwards found a couple of very useful extensions for the Firefox browser which made it very easy to connect to EC2 and S3 and to launch and manage instances. Overall, I think that it took me most of one day getting connected and figuring out how it all worked and perhaps another day after that to set up a machine with all of the software and tools that I required."*

*"Actually it was surprisingly easy to get up and running. However, we learned that we really needed to do a lot of exploration and benchmarking before we decide on an optimal configuration of Cloud resources."*

#### **D.5 Which parts of AWS have you been using in your work - eg EC2, S3, Elastic MapReduce, Cloudwatch**

- D.5.1 EC2 was used by all respondents. S3 and EBS were used by several, with one respondent making use of MapReduce. Cloudwatch was not used by any of the researchers, but one researcher currently using a proprietary monitoring system was considering it.

#### **D.6 How would you Describe your use of Cloud? (Data set, duration of computation, parallelism etc)**

- D.6.1 Respondents mainly used EC2 for processing datasets of ~10s of GBs in a relatively short space of time. No respondents made use of MPI or OpenMP parallelisation.

*"A typical night of data is several GB of raw CCD digital image data. We were particularly interested in having the ability to turn around a full reprocessing of all our data (about 500 nights worth, and that is expected to grow) on a 24-48 hour time-period. We found that if we could monopolize about 80 cores of EC2, each night taking around 3-4 hours to process (as we measured) then this would certainly be possible. Our data processing applications make no use of MPI or OpenMP parallelism: We are completely process-parallel in that we have a pipeline of long-running serial programs bundled together into a job for each night. The general model is a set of executable codes (Python, Perl, shell scripts, C, C++, Fortran) that are called from a main batch job script."*

## **D.7 Is there anything you would like to have done with AWS but could not?**

- D.7.1 One respondent wanted AWS to be “more like Google App Engine”, which of course begs the question as to why they were not using Google’s cloud service. Most other respondents answered ‘No’ – but added caveats. A common experience seemed to be that some re-thinking of their initial approach was required.

*“Greater ability to predict performance, eg by knowing whether other VMs were contending for resources.”*

*“Since we anticipated that we would have to experiment to find optimal solutions in EC2 ourselves, we were open to re-thinking certain things (but not everything) about our applications”*

*“No, I don't think so. I found it to be very flexible. Having used other computing resources (both local and national facilities) I found EC2 refreshing because I had root access to the machines so I could configure them and install software as I wanted. This was hugely beneficial - typically when I've used more traditional facilities I waste days of time trying to install newer versions of the usually outdated software on such systems as a user (ie without root access). I realize that some users may not want to have to configure things themselves, but personally I found this to be a highly efficient approach.”*

*“A better capability for low-latency cluster support would be great.”*

## **D.8 Were there any privacy or other regulatory/legal issues in your work?**

- D.8.1 None of the respondents had any regulatory or legal issues that affected their use of AWS. However, one respondent had encountered this issue in the past:

*“Previously, there was concern regarding potential HIPAA violations when using electronic medical data (previous project). Amazon allows for adequate security where a technical solution is feasible (see Amazon's white paper). Additionally, the use of encrypted datasets can alleviate privacy concerns (AES standard encryption).”*

## **D.9 Were there any other surprises. Pleasant or otherwise?**

- D.9.1 Several respondents described their ‘surprise’ at how easy the service was to use. However, nearly all had some issues to resolve – these were widely varied and seemed to be quirks relating the specifics of their individual requirements and usage profile and methods.

*“Amazon has been very responsive regarding funding and follow-up. We have enjoyed working with them.”*

*“One surprise was that we learned that our experiment's CVS server specifically blocked access from Amazon EC2, and we had to use an alternative means to actually check out our software onto the cloud!... Also, our initial selections of generic images to customize from the contributed public EC2 images were kind of problematic. We learned to locate the source of these images and understand who made them and why.”*

*“Mostly pleasant. Most tricky thing has been the OpenGL support which was hard to resolve.”*

*“My main surprise was at how easy it was to use. I found using the GUI interface provided by the ElasticFox Firefox extension to make it extremely easy to manage and control large numbers of instances. I don't recall any unpleasant surprises.”*

**D.10 Will you continue to pay to use a public cloud at the end of the Amazon funding?**

D.10.1 This question elicited a range of responses. Several respondents would like to continue to use AWS but were unsure of how to do this after their grant ran out. Other respondents felt that the costs of AWS were too high for non-grant funded work.

*"Potentially yes. I don't have need for large amounts of computing time on a regular basis, but if and when I next do I'll definitely look at using some form of Cloud Computing. I've been extolling its virtues to colleagues! One thing that is not clear to me (within the US funding system of course) is whether money from federal grants (which sometimes stipulate that the money cannot be used to buy computing equipment) can be used to buy Cloud services. I suspect that they can though. When we next need to upgrade local computers I will be very tempted to consider putting some of those funds into buying cloud time instead - it seems to be a very efficient use of the money."*

*"For my work, no. The expenses quickly accumulate, 10 c1.medium, for 10 hrs is \$17. I can foresee bills running into the hundreds of dollars."*

*"Unlikely. Our analysis of the cost means that unless a significant portion of the computing resources were donated by Amazon EC2 or paid for by our funding agency under some agreement, we would have a hard time justifying the cost. In particular the main cost is the data storage and transfers. The actual compute costs are less than a fifth of the overall costs. Moving data in and out of the Cloud and storing it there seems to be the major impediment to doing data driven science in the AWS environment. For us, a private Cloud run by the U.S. Department of Energy is likely to offer the advantages we were looking for without the large data movement and storage costs."*

*"I hope we can continue with the funding. We do not have that many resources for our project. I think it's great that Amazon is doing this for academia. Go Amazon!"*

This page is intentionally blank

## E Further details on current cloud providers

### E.1 Introduction

E.1.1 This annex contains some further information in support of Section 5. Any prices listed were correct as of April 2010.

### E.2 Storage as a Service

#### *Other providers*

- **iDrive:** 2GB free storage; monthly charges: <http://www.idrive.com/>
- **DropBox:** 2GB free storage; monthly charges: <http://www.dropbox.com>
- **iBackup:** monthly charges: <http://www.ibackup.com>
- **Box.Net:** 1GB free; monthly charges: <http://www.box.net/>
- **ElephantDrive:** 1GB free; monthly charges: <http://www.elephantdrive.com/>; *also makes use of S3.*

### E.3 Infrastructure as a Service

#### *Amazon AWS*

E.3.1 The table below provides key service information about the compute and storage services available under AWS:

<b>Compute</b>	<b><u>EC2</u></b>
Charges	Hourly; variable depending on O/S and availability zone
O/S	7 flavours of Linux, Open Solaris and Windows Server 2003/2008
Network charges	per GB transferred, except with other AWS services in the same availability zone
Backups	to S3 or EBS; per GB pricing
Lowest Cost Server Configuration	1.7 GB memory, 160 GB disk, 1 EC2 Compute Unit (1 virtual core equivalent to a 1.0-1.2 GHz 2007 Opteron or 2007 Xeon processor), 32-bit O/S
API	REST and SOAP
Associations	Various providers have set up AMIs that contain their Cloud software offerings as a baseline for other developments.

<b>Storage<sup>7</sup></b>	<b><u>S3<sup>8</sup></u></b>
Type	Web service oriented

<sup>7</sup> Virtual servers running under EC2 can make use of S3, EBS, and *ephemeral* storage: data space available on each machine instance up to the specified size of the image; lost when the virtual machine is terminated.

<sup>8</sup> Amazon S3 can be set up as a backup facility, for persistent storage for specific uses or applications; buckets can be configured with ACLs; payment against downloads via Amazon DevPay (Amazon takes a transaction fee).

Charges	per GB per month; per GB charges for data transfer and per 10,000 requests; vary by availability zone.
Content Delivery	Amazon CloudFront CDN
API	REST and SOAP
<b>Storage</b>	<b><u>EBS</u></b>
Type	SAN-like block store
Charges	per GB per month and per million requests
Content Delivery	n/a
API	n/a

### ***Rackspace***

- E.3.2 RackSpace provide dedicated servers, charged monthly, managed private Clouds, and Cloud Computing offerings with various charges applicable. Rackspace additionally provide hosted email, charged per mailbox, as either Rackspace's own variant (the acquired webmail.us) or Microsoft Exchange. RackSpace's Server Backup and Cloud Drive can provide for certain storage needs. RackSpace are one of the 37 organisations supporting the Open Cloud Manifesto. SliceHost, provider of monthly-charged Linux instances, has also become a RackSpace company, and JungleDisk is another RackSpace subsidiary offering monthly-charged storage.
- E.3.3 RackSpace has 8 data centre locations, with its primary data centre located in San Antonio, 4 other US locations, and both European data centres located in the UK. The 8<sup>th</sup> data centre is located in Hong Kong.
- E.3.4 RackSpace's listed Cloud offerings (<http://www.rackspacecloud.com>) include the following:

<b>Compute</b>	<b><u>Cloud Servers</u></b>
Charges	Hourly or Monthly
O/S	Linux only
Network charges	bandwidth charged per GB; outgoing almost 3 times price of incoming.
Backups	to Cloud Files; per GB pricing, also charges per 500 requests
Lowest Cost Server Configuration	256 MB memory, 10 GB disk
API	REST
Associations	RightScale built using Rackspace Cloud Servers API.
<b>Storage</b>	<b><u>Cloud Files</u></b>
Type	Web service oriented
Charges	per GB
Content Delivery	Limelight CDN
API	REST, Java, PHP, Ruby, Python and C#/.NET bindings

<b>Web Hosting (Platform)</b>	<b>Cloud Sites</b>
Supports	Windows and Linux; Content Management Systems such as Drupal; shopping software such as ZenCart
Charges	Monthly

### **Flexiscale**

1.1.2 Flexiscale public cloud is provided by Flexiant, whose other product offering, *Extility*, provides for Private Clouds and further demonstrates that any element of the computing stack can be packaged and sold over data centres in some way.

E.3.5 The Flexiscale public cloud offers:

<b>Compute</b>	
Charges	Based on non-refundable purchases of "units", in increments of 1,000
O/S	Linux (3 flavours) and Windows Server
Network charges	5 units per GB transferred
Backups	Snapshots; create your own off-site.
Lowest Cost Server Configuration	0.5GB memory, 1 CPU, costing 2 units/hour
API	SOAP
<b>Storage</b>	
Type	Virtual disks on a SAN
Charges	5 units per month per GB allocated – whether used or not; disk I/O charged at 2 units per GB transferred
Bandwidth	5 units per GB transferred
Content Delivery	n/a
API	REST; Java, PHP, Ruby, Python and C#/.NET bindings

### **Terremark vCloud Express**

E.3.6 Amongst services such as managed hosting and storage, Terremark offer two Cloud Computing products: vCloud Express (in Beta) and the contract-based Enterprise Cloud, both built above VMware. Currently there appears to be no separate storage service offering.

E.3.7 vCloud Express offers a similar pricing model to that of Amazon AWS though with a few variations:

<b>Compute</b>
----------------

Charges	Based on the number of virtual processors and memory footprint, plus a charge for the system disk. Applied when servers are configured, regardless of whether they are deployed for use or powered off.
O/S	Templates for 3 flavours of Linux and Windows Server 2003 and 2008. OS/2 and SCO Unixware 7 are listed options when creating a "Blank Server".
Network charges	per GB transferred
Backups	n/a
Lowest Cost Server Configuration	0.5GB memory, 1 virtual processor
API	REST; Python binding
Associations	n/a

- E.3.8 Terremark has 12 data centre locations, with 4 of these in London and Mainland Europe. A data centre is reported in Dallas, though this may be a partnership with Digital Realty Trust.

### **ElasticHosts**

- E.3.9 **ElasticHosts** provide configurable pre-paid hourly and monthly-charged servers. Servers can be configured and started manually or via HTTP calls. Currently there appears to be no separate storage service offering. A free 5-day trial of a certain number and configuration of virtual servers is currently offered with VNC access supported.

<b>Compute</b>	
Charges	Hourly or Monthly
O/S	4 flavours of Linux, OpenSolaris and Windows Server 2008
Network charges	per GB
Backups	User responsible
Lowest Cost Server Configuration	Hourly: 2GHz processor, 1.7GB memory, 160GB disk. Monthly: 2GHz processor, 1GB memory, 5GB disk, 10GB data transfer.
API	HTTP (ReST)
Associations	n/a

- 1.1.3 ElasticHosts currently uses 2 data centres "near London" – one in Maidenhead courtesy of BlueSquare Data, another courtesy of Peer1 in Fleet.

### **NewServers**

- E.3.10 NewServers provides non-virtualised infrastructure – "bare metal" - in hourly and monthly-charged units, using Dell servers. This increases the potential performance, but with a likely compromise in flexibility.

- E.3.11 NewServers operate from the NAP of Americas: Miami data centre, owned and operated by Terremark.

<b>Compute</b>	
Charges	Hourly or Monthly; \$20 deposit needed on account creation
O/S	Centos (64-bit) or Windows Server 2003; install Ubuntu, Solaris, Debian, RHEL and VmwareESX on request
Network charges	Includes 3GB bandwidth per hour; charges per GB above this
Backups	
Lowest Cost Server Configuration	Hourly: 2.8GHz processor, 1GB memory, 36GB disk.
API	XML over HTTPS
Associations	n/a

### **GoGrid**

- E.3.12 **GoGrid** offer dedicated servers, charged monthly or annually, and Cloud Servers in two flavours of Linux and with Windows versions, with monthly charging for storage.
- E.3.13 Large datasets can be hosted in GoGrid's Cloud Storage for a one-off charge based on a shipped external hard drive. There will be a clear trade-off between the cost incurred for this one transfer and the speed and cost of using network transfer. Typically, data can be transitioned to GoGrid using typical mechanisms of SCP, FTP, SAMBA/CIFS, and RSYNC, and provided over the GoGrid CDN. The first 10GB of storage is provided free.
- E.3.14 GoGrid has a single data centre in the San Francisco Telecom Center, a facility shared with AT&T and MCI and through which emergency calls are routed, entailing preventive measures against seismic activities.

### **Joyent**

- E.3.15 **Joyent** offers a monthly-charged infrastructure, as well as a web development PaaS. The monthly charges include an initial quantity of data transfer (10TB) and various monthly annual fees apply to other aspects of provision.
- E.3.16 Joyent has been using Sun systems, with virtualised platform-oriented servers, called Joyent Accelerators, built over OpenSolaris. Joyent was also partnered with Sun to offer free hosting for a tranche of Facebook developers. Potential users may wonder about continued support for such activities from Sun following the Oracle takeover. Previously, Twitter had used Joyent before their move away to NTT America.

### **Other providers:**

E.3.17 For comparability purposes, offerings from the likes of Carpathia Hosting Inc,<sup>9</sup> Layered Technologies,<sup>10</sup> 3Tera<sup>11</sup> and Skytap<sup>12</sup> should also be examined.

## **E.4 Software as a Service (SaaS)**

E.4.1 Although this is not the specific focus of this report, SaaS is the most mature service model, and certain SaaS applications may be of use within research collaborations. While avoiding significant depth of coverage in this section, an overview of possible SaaS offerings is given.

### **Email and Documents**

E.4.2 Email and office software are widely in use and have relatively well understood compute requirements. Additionally, such applications can be used with relative immediacy on the web, in stark contrast to the need to install, configure, and maintain such software on an individual machine. Google Mail and Google Apps provide for such services and offer web-based collaborative authoring and a limited amount of document management. Certain organisations, and a number of UK Universities, have contracted with Google for the direct provision of these services<sup>13</sup>. Alternatives are readily available and include: Adobe Buzzword,<sup>14</sup> Peepel WebWriter,<sup>15</sup> WebEx,<sup>16</sup> Zoho Office<sup>17</sup> alongside Yahoo Mail, Windows Live (Hotmail), and so on.

### **Other SaaS applications**

E.4.3 Miller<sup>18</sup> provides coverage of a broad range of SaaS applications. A subset of these is summarised from this work in the table below.

<b>Application</b>	<b>Examples</b>
Calendar	Google, Yahoo, Windows Live, CalendarHub, Hunt Calendars, Calendar Net
Schedules	Diarised, Windows Live Events, Schedulebook, AppointmentQuest
Planning / Task Management	Bla-bla List, Hiveminder, Remember the Milk, Tudu List, HiTask, Zoho Planner
Event Management	Conference.com, RegOnline, Event Wax
Project Management	BaseCamp, Project Drive, Zoho Projects, onProject
Web Databases	Zoho Creator / Zoho DB & Reports, MyWebDB, Cebase, QuickBase, Lazybase
Bookmarking	BlinkList, Clipmarks, del.icio.us, Tagseasy

<sup>9</sup> <<http://www.carpathiahosting.com/>> [accessed 21 April 2010]

<sup>10</sup> <<http://www.layeredtech.com/>> [accessed 21 April 2010]

<sup>11</sup> <<http://www.3tera.com/>> [accessed 21 April 2010]

<sup>12</sup> <<http://www.skytap.com/>> [accessed 21 April 2010]

<sup>13</sup> See: <<http://www.google.com/a/help/intl/en/edu/customers.html>> [accessed 21 April 2010] for specific examples.

<sup>14</sup> <<http://buzzword.acrobat.com/>> [accessed 21 April 2010]

<sup>15</sup> <<http://www.peepel.com/>> [accessed 21 April 2010]

<sup>16</sup> <<http://www.weboffice.com/>> [accessed 21 April 2010]

<sup>17</sup> <<http://office.zoho.com/>> [accessed 21 April 2010]

<sup>18</sup> Miller, M. (2009) *Cloud Computing: Web-Based Applications that Change the Way you Work and Collaborate Online*, Que Publishing.

Photo Editing	FotoFlexer, Preloadr, Snipshot
Photo Sharing	dotPhoto, Flickr, Photobucket, Picasa Web Albums
Desktops	ajaxWindows, eyeOS, g.ho.st, YouOS
Web Conferencing	Genesys Meeting Center, IBM Lotus Sametime, Microsoft Office Live Meeting, WebEx, Zoho Meeting
Groupware	Contact Office, Google Sites, Project Spaces, teamspace
Blogs and Wikis	Blogger, TypePad, WordPress, Pbwiki, wikihost.org, Wikispaces, Zoho Wiki

E.4.4 These SaaS will all have different terms and conditions, pricing and support, and comparative review is necessary.

## E.5 Derivative Services

### *CloudKick*

E.5.1 CloudKick<sup>19</sup> has similarities to RightScale in that multiple Clouds can be managed through a single web interface. Seven cloud providers, including EC2, Rackspace, and GoGrid are supported. However, while servers can be launched on these providers using the CloudKick web interface, there is no support for configuration; CloudKick provides capability for launching and monitoring, but does not yet have anywhere near as wide a feature set as RightScale.

E.5.2 One useful spin-off from CloudKick is the Open Source libcloud library, written in Python, that provides hooks into the various Cloud providers.

### *Elastra*

E.5.3 Elastra<sup>20</sup> provides software for architecting and deploying Cloud systems. The Elastra AWS deployment software is itself bundled into an Amazon AMI, requiring deployment on EC2 prior to use.

## E.6 Platform as a Service

### *Google App Engine*

E.6.1 An application currently gets 500 MB of persistent storage and CPU and bandwidth for about 5 million page views a month – this would represent a reasonably popular application. Beyond these, billing applies (upper limits can be imposed); as at February 2010:

<sup>19</sup> <<http://incubator.apache.org/libcloud/about.html>> [accessed 21 April 2010]

<sup>20</sup> <<http://www.elastra.com>> [accessed 21 April 2010]

<b>Resource</b>	<b>Unit</b>	<b>Unit cost</b>
Outgoing Bandwidth	Gigabytes	\$0.12
Incoming Bandwidth	Gigabytes	\$0.10
CPU Time	CPU hours	\$0.10
Stored Data	gigabytes per month	\$0.15
Recipients Emailed	recipients	\$0.0001

### ***Force.com***

- E.6.2 Force.com is the Platform underlying Salesforce.com that allows users to build their own customised versions of applications for Salesforce.com within the same multi-tenant architecture. Applications built with this Platform can be sold via AppExchange. Force.com provides a number of APIs enabling connections to Google App Engine, Facebook, SAP R/3, Oracle Financials and other applications.

## F Cloud Computing Research

F.1.1 There have been relatively few academic conferences and workshops held regarding Cloud Computing, though there are certainly signs of growth.

- First International Conference on Cloud Computing Technology (CloudCom 2009), Beijing, China, December 1-4, 2009; proceedings available from Springer<sup>21</sup>
- The 10th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing<sup>22</sup> (CCGrid) 2010, May 17-20, 2010, Melbourne, Victoria, Australia, including the 2nd International Symposium on Cloud Computing (Cloud 2010)<sup>23</sup>; also, the IEEE International Symposium on Cluster Computing and the Grid (CCGrid) 2010, Shanghai, China which incorporated the International Workshop on Cloud Computing (Cloud 2009).
- Cloud-based Services and Applications Workshop<sup>24</sup> at the 5th IEEE International Conference on e-Science, Oxford, UK, December 9-11 2009
- IGT 2008 World Summit of Cloud Computing<sup>25</sup>, Israel, December 1-2 2008

F.1.2 Research published in these avenues can be readily related to the cloud stack, reprised below:

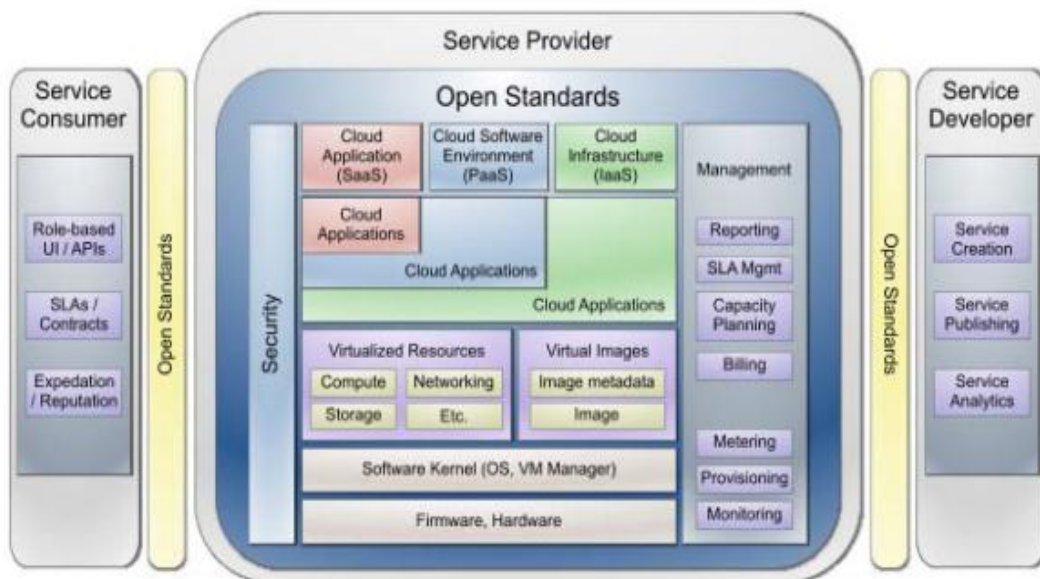


Figure F-1: the cloud stack. From Cloud Computing Use Cases White Paper, version 2.0<sup>26</sup>

21 See: <<http://www.springer.com/computer/communication+networks/book/978-3-642-10664-4>> [Accessed 21 April 2010]

22 <<http://www.manjrasoft.com/ccgrid2010/>> [Accessed 21 April 2010]

23 <<http://www.cloudbus.org/cloud2010/Welcome.html>> [Accessed 21 April 2010]

24 <<http://www.oerc.ox.ac.uk/ieee/workshops/cloud>> [Accessed 21 April 2010]

25 <<http://www.cloudcomputing.org.il/>> [Accessed 21 April 2010]

26 Available via <[http://cloud-computing-use-cases.googlegroups.com/web/Cloud\\_Computing\\_Use\\_Cases\\_Whitepaper-2\\_0.pdf](http://cloud-computing-use-cases.googlegroups.com/web/Cloud_Computing_Use_Cases_Whitepaper-2_0.pdf)> [Accessed 21 April 2010]

- F.1.3 The following provides examples of academic research that is either published, or scheduled for publication, as a starting point for those interested in Cloud Research per se. The papers listed have been broadly classified according to the above figure, however this classification is for illustrative purposes only and publications may cross into many more areas of the cloud stack.

#### ***Service Consumer (SLAs / Contracts)***

- Mathias Dalheimer and Franz-Josef Pfreundt, (2009). "**GenLM: License Management for Grid and Cloud Computing Environments**". In Proceedings of the IEEE/ACM International Symposium on Cluster Computing and the Grid, 2009. IEEE Press.

#### ***Service Consumer (Expediation / Reputation)***

- Wenjuan Li and Lingdi Ping, (2009) "**Trust Model to Enhance Security and Interoperability of Cloud Environment**". In Proceedings of First International Conference, CloudCom 2009, Beijing, China, December 1-4, 2009. LNCS 5931. Springer.

#### ***Open Standards***

- Anand Govindarajan and Lakshmanan G, (2010). "**Overview of Cloud Standards**". In Nikolaos Antonopoulos and Lee Gillam (eds.), *Cloud Computing: Principles, Systems and Applications*. Springer. ISBN: 978-1-84996-240-7

#### ***Security***

- Wenjuan Li and Lingdi Ping (2009) "**Trust Model to Enhance Security and Interoperability of Cloud Environment**". In Proceedings of First International Conference, CloudCom 2009, Beijing, China, December 1-4, 2009. LNCS 5931. Springer.
- Hongwei Li, Yuanshun Dai, Ling Tian, and Haomiao Yang (2009) "**Identity-Based Authentication for Cloud Computing**". In Proceedings of First International Conference, CloudCom 2009, Beijing, China, December 1-4, 2009. LNCS 5931. Springer.
- Sadie Creese, Paul Hopkins, Siani Pearson, and Yun Shen, (2009), "**Data Protection-Aware Design for Cloud Services**". In Proceedings of First International Conference, CloudCom 2009, Beijing, China, December 1-4, 2009. LNCS 5931. Springer.

#### ***Cloud Application (SaaS)***

- Jinyu Han, Min Hu, and Hongwei Sun, (2009). "**Search Engine Prototype System Based on Cloud Computing**". In Proceedings of First International Conference, CloudCom 2009, Beijing, China, December 1-4, 2009. LNCS 5931. Springer.

#### ***Cloud Infrastructure (IaaS)***

- Daniel Nurmi, Rich Wolski, Chris Grzegorzczak, Graziano Obertelli, Sunil Soman, Lamia Youseff, Dmitrii Zagorodnov, (2009). "**The Eucalyptus Open-source Cloud-computing System**". In Proceedings of the IEEE/ACM International Symposium on Cluster Computing and the Grid, 2009. IEEE Press.

- Christian Baun, Marcel Kunze, (2009). "**Building a Private Cloud with Eucalyptus**". In proceedings of Cloud-based Services and Applications Workshop at the 5th IEEE International Conference on e-Science, Oxford, UK, December 9-11 2009. IEEE Press.

### **Cloud Software Environment (PaaS)**

- Fabrizio Marozzo, Domenico Talia and Paolo Trunfio, (2010). "**A Peer-to-Peer Framework for Supporting MapReduce Applications in Dynamic Cloud Environments**". In Nikolaos Antonopoulos and Lee Gillam (eds.), *Cloud Computing: Principles, Systems and Applications*. Springer. ISBN: 978-1-84996-240-7

### **Virtualized Resources (Networking)**

- Francesco Pamieri and Silvio Pardi, (2010). "**Enhanced network support for scalable computing clouds**". In Nikolaos Antonopoulos and Lee Gillam (eds.), *Cloud Computing: Principles, Systems and Applications*. Springer. ISBN: 978-1-84996-240-7

### **Virtualized Resources (Storage)**

- Steve Todd, Dan Hushon, (2009) "**Scientific Lineage and Object-Based Storage Systems**". In proceedings of Cloud-based Services and Applications Workshop at the 5th IEEE International Conference on e-Science, Oxford, UK, December 9-11 2009. IEEE Press.
- Xiaoming Gao, Mike Lowe, Yu Ma, and Marlon Pierce, (2009). "**Supporting Cloud Computing with the Virtual Block Store System**". In proceedings of Cloud-based Services and Applications Workshop at the 5th IEEE International Conference on e-Science, Oxford, UK, December 9-11 2009. IEEE Press.

### **Virtual Images (Image)**

- William Voorsluys, James Broberg, Srikumar Venugopal, and Rajkumar Buyya, (2009). "**Cost of Virtual Machine Live Migration in Clouds: A Performance Evaluation**". In Proceedings of First International Conference, CloudCom 2009, Beijing, China, December 1-4, 2009. LNCS 5931. Springer.
- Kyrre Begnum, Nii Apleh Lartey, and Lu Xing, (2009). "**Cloud-Oriented Virtual Machine Management with MLN**". In Proceedings of First International Conference, CloudCom 2009, Beijing, China, December 1-4, 2009. LNCS 5931. Springer.

### **Management (Provisioning)**

- Ying Song, Hui Wang, Yaqiong Li, Binquan Feng, Yuzhong Sun (2009), "**Multi-Tiered On-Demand Resource Scheduling for VM-Based Data Center**". In Proceedings of the IEEE/ACM International Symposium on Cluster Computing and the Grid, 2009. IEEE Press.
- Takahiro Hirofuchi , Hirotaka Ogawa , Hidemoto Nakada, Satoshi Itoh and Satoshi Sekiguchi, (2009). "**A Live Storage Migration Mechanism over WAN for Relocatable Virtual Machine Services on Clouds**". In Proceedings of the IEEE/ACM International Symposium on Cluster Computing and the Grid, 2009. IEEE Press.

### **Management (SLA Mgmt)**

- Bin Li and Lee Gillam (2009) "**Towards Job-Specific Service Service Level Agreements in the Cloud**". In proceedings of Cloud-based Services and Applications Workshop at the 5th IEEE International Conference on e-Science, Oxford, UK, December 9-11 2009. IEEE Press.
- Waheed Iqbal, Matthew Dailey, and David Carrera (2009), "**SLA-Driven Adaptive Resource Management for Web Applications on a Heterogeneous Compute Cloud**". In Proceedings of First International Conference, CloudCom 2009, Beijing, China, December 1-4, 2009. LNCS 5931. Springer.
- Rajkumar Buyya, Suraj Pandey, and Christian Vecchiola (2009), "**Cloudbus Toolkit for Market-Oriented Cloud Computing**". In Proceedings of First International Conference, CloudCom 2009, Beijing, China, December 1-4, 2009. LNCS 5931. Springer.

### **Management (Billing)**

- Asoke K Talukder, Lawrence Zimmerman, and Prahalad H.A, (2010). "**Cloud Economics: Principles, Costs and Benefits**". In Nikolaos Antonopoulos and Lee Gillam (eds.), *Cloud Computing: Principles, Systems and Applications*. Springer. ISBN: 978-1-84996-240-7

### **Management (Provisioning)**

- Yuanshun Dai, Yanping Xiang, and Gewei Zhang (2009) "**Self-healing and Hybrid Diagnosis in Cloud Computing**". In Proceedings of First International Conference, CloudCom 2009, Beijing, China, December 1-4, 2009. LNCS 5931. Springer.
- Yonggang Wang, Sheng Wang, and Daliang Zhou, (2009). "**Retrieving and Indexing Spatial Data in the Cloud Computing Environment**". In Proceedings of First International Conference, CloudCom 2009, Beijing, China, December 1-4, 2009. LNCS 5931. Springer.
- Rajiv Ranjan, Liang Zhao, Xiaomin Wu, Anna Liu, Andres Quiroz and Manish Parashar, (2010). "**Peer-to-Peer Cloud Provisioning: Service Discovery and Load-Balancing**". In Nikolaos Antonopoulos and Lee Gillam (eds.), *Cloud Computing: Principles, Systems and Applications*. Springer. ISBN: 978-1-84996-240-7

### **Management (Monitoring)**

- Milan Milenkovic, Enrique Castro-Leon, and James R. Blakley, (2009). "**Power-Aware Management in Cloud Data Centers**". In Proceedings of First International Conference, CloudCom 2009, Beijing, China, December 1-4, 2009. LNCS 5931. Springer.

### **Service Developer (Service Creation)**

- Geoffrey Fox, Xiaohong Qiu, Scott Beason, Jong Choi, Jaliya Ekanayake, Thilina Gunarathne, Mina Rho, Haixu Tang, Neil Devadasan, and Gilbert Liu (2009), "**Biomedical Case Studies in Data Intensive Computing**". In Proceedings of First International Conference, CloudCom 2009, Beijing, China, December 1-4, 2009. LNCS 5931. Springer.
- Chunming Rong (2009), "**An Industrial Cloud: Integrated Operations in Oil and Gas in the Norwegian Continental Shelf**". In Proceedings of First International Conference, CloudCom 2009, Beijing, China, December 1-4, 2009. LNCS 5931. Springer.

- Hyun Jung La and Soo Dong Kim, (2009). "**A Systematic Process for Developing High Quality SaaS Cloud Services**". In Proceedings of First International Conference, CloudCom 2009, Beijing, China, December 1-4, 2009. LNCS 5931. Springer.

This page is intentionally blank

## G Cloud-based Datasets

- G.1.1 Amazon hosts a number of datasets on disk volumes (EBS) for research purposes.<sup>27</sup> These include Federal Reserve Economic Data, Daily Global Weather Measurements, Wikipedia Page Traffic Statistics, and an OpenStreetMap Rendering Database. Three reasonably large datasets are the Sloan Digital Sky Survey DR6 Subset, a Human Genome Dataset, and Wikipedia in XML.

### Sloan Digital Sky Survey DR6 Subset

 Printer Friendly  Save to del.icio.us

The Sloan Digital Sky Survey is the most ambitious astronomical survey ever undertaken. [Discussion](#)

[Reviews](#)

<b>Submitted By:</b>	thakar
<b>US Snapshot ID (Linux/Unix):</b>	snap-3740f35e
<b>US snapshot ID (Windows):</b>	snap-1047f479
<b>Size:</b>	180 GB
<b>Creation Date:</b>	09/28/2009
<b>Last Updated:</b>	09/28/2009
<b>License:</b>	Creative Commons: Attribution
<b>Source:</b>	Sloan Digital Sky Survey

\*Note: The data for this data set is in an MDF SQL Server file. The data on the Linux snapshot would have to be converted to a native Linux database format.

The Sloan Digital Sky Survey is the most ambitious astronomical survey ever undertaken. The survey has mapped one-quarter of the entire sky in detail, determining the positions and absolute brightnesses of hundreds of millions of celestial objects. It has also measured the distances (redshifts) to more than a million galaxies and quasars. This is a small (~ 5%) subset of the Sloan Digital Sky Survey Data Release 6 Catalog Archive. The subset was generated by extracting all the data in a small region of the sky from the SDSS DR6 "best" dataset (BestDR6) which was released to the public on June 29, 2007.

*Figure G-1: Information from the Amazon website about the Sloan Digital Sky Survey dataset.*

<sup>27</sup> <<http://aws.amazon.com/publicdatasets/>> [accessed 21 April 2010]

## Illumina - Jay Flatley (CEO of Illumina) Human Genome Data Set

Printer Friendly Save to del.icio.us

Jay Flatley (CEO of Illumina) human genome data set.

Discussion

Reviews

Submitted By:	Dave@AWS
US Snapshot ID (Linux/Unix):	snap-53b3cb3a
US snapshot ID (Windows):	snap-25b3cb4c
Size:	350 GB
Creation Date:	01/12/2009
Last Updated:	01/12/2009
License:	Creative Commons: Attribution Share Alike
Source:	Illumina

This data set contains the raw export files of the first genome sequenced by [Illumina Individual Genome Service](#) using Illumina's Genome Analyzer technology of paired 75-base reads. 92,254,659,274 bases were used to generate a consensus sequence with coverage of 32x average depth. The genome was obtained via peripheral blood of Jay Flatley, CEO of Illumina.

Figure G-2: Information from the Amazon website about the Human Genome dataset

## Wikipedia XML Data

Printer Friendly Save to del.icio.us

A complete copy of all Wikimedia wikis, in the form of wikitext source and metadata embedded in XML.

Discussion

Reviews

Submitted By:	Santiago@AWS
US Snapshot ID (Linux/Unix):	snap-8041f2e9
Size:	500GB
Creation Date:	08/07/2009
Last Updated:	08/07/2009
License:	Creative Commons: Attribution Share Alike
Source:	Wikimedia Foundation ( <a href="http://download.wikipedia.org/backup-index.html">http://download.wikipedia.org/backup-index.html</a> )

This data set contains a complete copy of all Wikimedia wikis, in the form of wikitext source and metadata embedded in XML as provided by the [Wikimedia Foundation](#).

The data set will be updated every month and the 3 previous months will always be available for use. We will list previous snapshots in the text of this description.

Figure G-3: Information from the Amazon website about the Wikipedia XML dataset