Cloud Storage and Bioinformatics in a Private Cloud Deployment: Lessons for Data Intensive Research

Victor Chang[1,2], Robert John Walters[1], and Gary Wills[1]

[1] Electronics and Computer Science, University of Southampton, Southampton SO 17 1BJ, U.K.
[2] School of Computing and Creative Technologies, Leeds Metropolitan University, Headingley, Leeds LS6 3QS, U.K.
{vic1e09,rjw1,gbw}@ecs.soton.ac.uk, V.I.Chang@leedsmet.ac.uk

Abstract. This paper describes service portability for a private cloud deployment, including a detailed case study about Cloud Storage and bioinformatics services developed as part of the Cloud Computing Adoption Framework (CCAF). Our Cloud Storage design and deployment is based on Storage Area Network (SAN) technologies, details of which include functionalities, technical implementation, architecture and user support. Experiments for data services (backup automation, data recovery and data migration) are performed and results confirm backup automation is completed swiftly and is reliable for data-intensive research. The data recovery result confirms that execution time is in proportion to quantity of recovered data, but the failure rate increases in an exponential manner. The data migration result confirms execution time is in proportion to disk volume of migrated data, but again the failure rate increases in an exponential manner. In addition, benefits of CCAF are illustrated using several bioinformatics examples such as tumour modelling, brain imaging, insulin molecules and simulations for medical training. Our Cloud Storage solution described here offers cost reduction, time-saving and user friendliness.

1 Introduction

Cloud Computing offers a variety of benefits including cost-saving, agility, efficiency, resource consolidation, business opportunities and Green IT [9-13, 16-18, 20, 23]. As more organisations adopt Cloud, the need for a standard, or a framework to manage both operation management and IT services is emerging. This framework needs to provide the structure necessary to ensure any Cloud implementation meets the business needs of Industry and Academia and include recommendations of best practices which can be adapted for different domains and platforms. Our framework is called the Cloud Computing Adoption Framework (CCAF). It helps organisations to achieve good Cloud design, implementation and services [11-20]. CCAF may be used from service strategy to design, development, test and user support stages. The CCAF seeks to address two problems in particular:

I. Ivanov et al. (Eds.): CLOSER 2012, CCIS 367, pp. 245–264, 2013.

- Calculating Cloud Business Performance systematically and coherently.
- Portability of services into the Cloud

This paper focuses on service portability which is the term we use to describe a recommended approach to Cloud adoption. Cloud adoption plays an important role in having a smooth transition to the Cloud environment. Beaty et al. [3] and Chang et al. [11,18,20] identify portability as an adoption challenge for organisational Cloud adoption. Although it is domain specific as there are different requirements for portability in each domain, communication between different types of clouds supplied by different vendors can be difficult to implement. Often work-arounds are needed which entail writing additional layers of APIs, or an interface or portal [2,3].

Service portability (portability in short) is illustrated using examples from Cloud Storage projects in the Healthcare industry where portability is influential in migrating existing infrastructure, platforms and applications to the Cloud and later developing new applications and services. The storage is provided using in-house private clouds, initially to provide a working IaaS infrastructure for medical databases, images and analysis in a secure and collaborative environment. These Cloud projects have been successfully delivered and provide a high level of user satisfaction and were followed up with further work to upgrade from IaaS to PaaS, which allows greater benefits, including better efficiency and better management of resources. We also present results from experiments for data services (backup automation, data recovery and data migration) which can help us to meet issues and challenges of data-intensive research. The structure of this paper is as follows. Section 2 describes the overview of Cloud Storage and Section 3 presents its deployment architecture and user support. Section 4 explains bioinformatics and its associated results. Section 5 discusses performance results for data-intensive storage. Section 6 presents topics of discussion and Section 7 sums up Conclusion and future work.

2 Healthcare Cloud Storage

Supported by NHS UK, Guy's and St Thomas NHS Trust (GSTT) and King's College London (KCL) have worked together on projects to implement Cloud Storage and deliver it as a service. The initial effort was directed to an evaluation of the technology and developed a proof of concept service. CCAF is instrumental and influential in the way Cloud Storage has been developed:

- Healthcare Cloud Storage is a PaaS system, and needed careful planning and a thorough implementation. This required integrated adoption of multiple vendors' solutions.
- Healthcare Cloud Storage is an area to experience rapid growth in user requirements and disk space consumption. Therefore, it had to be easy to use, and able to cope with increasing demand.
- Healthcare Cloud Storage is a new concept and implementation in the Health domain where private and in-house storage has been used in the past. Maintenance of data protection and security is a challenge.

Better performance in from Healthcare Cloud Storage than previous storage service is regarded as a benchmark and measurement for success by executives. Recommendations, strategy and support from CCAF provided useful good services. Healthcare Cloud Storage has used trials during its design and implementation to ensure it meets its requirement to provide a robust service.

Healthcare Cloud Storage is used by the Breast Cancer project. Breast cancer is the most common cancer in women and has a worldwide annual incidence of over 1 million cases. There are many thousands of data about patients (medical records) and tumours (detailed descriptions and images, and its relations to the patients). Data growth is rapid and the data needs to be carefully used and protected. The work involves integrating software and cloud technologies from commercial vendors including Oracle, VMWare, EMC, Iomega and HP. This is to ensure a solid infrastructure and platform is available. Researchers also use third party applications to access, view and edit tumour images from trusted locations. Security is enforced in terms of data encryption, SSL and firewalls. Ion addition to Cloud Storage, the Health Cloud platform also provides Bioinformatics services, which provide scientific visualisation and modelling of genes, proteins, DNA, tumour and brain images. Users are very supportive in this project and some of them use it daily.

2.1 Benefits from Adopting CCAF

Adopting CCAF assists with understanding of requirements, technical knowledge, use cases and issues to be aware of, before and during the project development. Healthcare Cloud Storage is implemented as a Private Cloud project and is divided into four stages summed up as follows.

Stage 1	Explore available technologies, understanding strength and weaknesses for each key technology. Capture user requirements to get into technical plans.
Stage 2	Propose a framework based on the outcomes in Stage 1 and CCAF, and carry out plans for building and validating the framework.
Stage 3	Propose and implement service oriented architecture for Cloud Storage based on CCAF. Offer services for users and research groups.
Stage 4	Continue service improvements and further integration with other services and other new requirements.

Healthcare Cloud offers a wide range of self- and automated services across secure networks. It has two different focuses. It must be easy to use and support several research groups (both synchronously and asynchronously) and be able to cope with frequent changes, updates and user activities. It must also be highly robust and stable, allowing data to be kept safe, secure and active for extended periods of time (ten years and above). Both aspects demand for the following four requirements:

- Automated backup.
- Data recovery and emergency services. Snapshots or disaster recovery are used.

- Quality of services: high availability, reliability and great usability.
- Security.

This needs the state-of-the-art design and implementation that the CCAF can offer. The CCAF positively influences the way the backup and storage are designed and deployed. CCAF also offers implementation insights such as integration, as it is a challenge to co-ordinate and to combine different research activities and repositories into a distributed storage. This leads to the use of third party applications and services to improve on the quality of services.

2.2 A Storage Area Network Made Up of Different Clusters of Network Attached Storage (NAS)

The Architecture design chosen uses two concurrent platforms. The first is based on Network Attached Storage (NAS), and the second is based on the Storage Area Network (SAN). The NAS platform provides great usability and accessibility for users. Each NAS may be allocated to a research group and operate independently. Then all the NAS can be joined up to establish a SAN. NAS supports individual backups with manual and automated options. One option is similar to the Dropbox pattern of backup enabling users to copy their files onto their allocated disk space without difficulty providing a backup facility which is easy to use and user-friendly. Such a manual service allows users to backup their resources onto a selected destination and can offer both compressed and uncompressed versions of backup as well as data encryption to enforce security.

The Storage Area Network (SAN) is a dedicated and extremely reliable backup solution offering a highly robust and stable platform. SAN can consolidate an organisational backup platform and can improve capabilities and performance of Cloud Storage. SAN allows data to be kept safe and archived for a long period of time, and is a chosen technology. A SAN can be made up of different NAS, so that each NAS can focus on a particular function.

The design of SAN focuses on SCSI, which offers dual controllers and dual networking gigabyte channels. Each SAN server is built on RAID system. RAID 10 is a good choice since it can boost the performance like RAID 0 but also has mirroring capability like RAID1. A SAN can be built to have 12TB of disk space, and a group of SAN can form a solid cluster, or a dedicated Wide Area of Network. There are written and upgraded applications in each SAN to achieve the following functions:

- Performance improvement and monitoring: This allows tracking the overall and specific performance of the SAN cluster, and also enhances group or individual performance if necessary.
- Disk management: When a pool of SAN is established, it is important to know which hard disks in the SAN serve for which servers or which user groups.
- Advanced backup: Similar functionalities to those described in the NAS, such as automation, data recovery and quality of services, are available here. The difference is more sophisticated techniques and mechanisms (use of enterprise software is optional) are required.

Some applications mainly based on PHP, MySQL and Apache have been written, to allow researchers to access the digital repository containing tumours. Users can access their Cloud Storage via browsers from trusted offices, and they need not worry about complexity, and work as if on their familiar systems. This Healthcare PaaS is a demonstration of enterprise portability. In addition, several upgrades have taken place to ensure the standard of Cloud Storage and quality of services. One example is the use of SSL certificates and the enforced authentication and authorisation of every user to improve on security. There is an automated service to backup important resources.

3 Healthcare Cloud Storage Deployment Architecture and User Support

This section describes how Cloud Storage is set up, and how its key functionality offers services and user support. Cloud Storage is a private-cloud SAN architecture made up of different NAS services, where each NAS is dedicated for one specific function. Design and Deployment is based on group requirements and their research focus.

3.1 Design and Deployment to Meet Challenges for Data Intensive Research

Design and deployment should meet challenges for data-intensive research challenges. Moore et al [25] and Bryant [4] point out that data-intensive research should meet demands for data recovery and data migration and allows a large number of data to be recovered and moved quickly and efficiently in ordinary operations and in emergency. This is suitable for Cloud Storage as the design and deployment must provide resilient, swift and effective services. Vo, Chen and Ooi [27] present their perspective on Cloud Storage and demonstrate how to perform experiments in data intensive environments, including performing read, write and transaction operations. They demonstrate their solution for data migration but there is a lack of consideration of data recovery which is important in the event of possible data loss. Abu-Libdeh, Princehouse and Weatherspoon [1] demonstrate their Cloud Storage case study which presents how “Failure Recovery” can get large-scaled data recovery and data migration completed. Although they demonstrate data migration and data recovery over months in their in-house development, they do not show the execution time for each data migration and recovery. This is an important aspect in Cloud Storage to allow each operation of large-scale data recovery and data migration to run smoothly and effectively. Design and deployment of Cloud Storage must meet demands in large-scaled backup automation, data recovery and data migration.

3.2 Selections of Technology Solutions

Selections of Technology Solutions are essential for Cloud Storage development as presented in Table 1.

Table 1. Selections of Technology Solutions

Technology selections	What is it used	Vendors involved	Focus or rationale	Benefits or impacts
Network Attached Storage (NAS)	To store data and perform automated and manual/ personal backup.	Iomega/EMC Lacie Western Digital HP	They have a different focus and set up. HP is more robust but more time-consuming to configure. The rest is distributed between RAID 0, 1 and 5.	Each specific function is assigned with each NAS. There are 5 NAS at GSTT/KCL site and 3 at Data Centre, including 2 for Archiving. Deployment Architecture is shown in Figure 1.
Infrastructure (networking and hosting solution)	Collaborator and in-house	University of London Data Centre	Some services need a more secure and reliable place. University of London Data Centre offers 24/7 services with around 500 servers in place, and is ideal for hosting solution.	Amount of work is reduced for maintenance of the entire infrastructure. It stores crucial data and used for archiving, which backup historical data and backup the most important data automatically and periodically.
Backup applications	Third party and in-house	Open Source Oracle HP Vmware Symantec In-house development	There is a mixture of in-house development and third party solution. HP software is used for high availability and reliability. The rest is to support backup in between NAS. Vmware is used for virtual storage and backup.	Some applications are good in a particular service, and it is important to identify the most suitable application for particular services.
Virtualisation	Third party	VMware VSphere and Citrix	It consolidates IaaS and PaaS in private cloud deployment.	Resources can be virtualised and saves effort such as replication.
Security	Third party and in-house	KCL/GSTT Macafee Symantec F5	Security is based on the in-house solution and vendor solution is focused on secure firewall and anti-virus.	Remote access is given to a list of approved users.

3.3 Deployment Architecture

There are two sites for hosting data, one is jointly at GSTT and KCL premises distributed in dedicated server rooms and the other is at University of London Data Centre to store and backup the most important data. Figure 1 shows the Deployment Architecture.

There are five NAS at GSTT and KCL premises and each NAS is provided for a specific function. Bioinformatics Group has the most demands. NAS 1 is used for their secure backup, and NAS 2 is used for their computational backup, which is then connected to Bioinformatics services. NAS 3 is used as an important gateway for backup and archiving and is an active service connecting with the rest. NAS 3 is shared and used by Cancer Epidemiology and BCBG Group. NAS 4 provides mirror services for different locations and offers an alternative in case of data loss. NAS 5 is initially used by Digital Cancer cluster, and helps to back up important files in NAS 3. There are two digital cancer clusters, which can back up between each other, and important data are backed up to NAS 8 for reliability and NAS 5 for local version. The reason for this is that a disaster recovery activity which took place in 2010 took two weeks full time to retrieve and recover data. Multiple backups ensure if one dataset is lost, the most recent archive (done daily) can be replaced without much time spent.

There are three NAS at the University of London Computing (Data) Centre (ULCC) where there are about 500 servers hosted for Cloud and HPC services. NAS 6 is used as a central backup database to store and archive experimental data and images. The other two advanced servers are customised to work as NAS 7 and 8 to store and archive valuable data. Performance for backup and archiving services is excellent and most data can be backed up in a short and acceptable time frame of less than one hour to back up data and images. This outcome is widely supported by users and executives. There are additional five high performance computing services based on Cloud technologies: Two are computational statistics to analyse complex data. The third one is a database to store confidential data and the fourth is on bioinformatics to help bioinformatics research. The last one is a virtualisation service that allows all data and backup to be in virtual storage format. These five services are not included in Cloud Storage for this paper.

3.4 User Support

The entire Cloud Storage Service has automated capability and is easy to use. This service has been in use without the presence of Chief Architect for six months, without major problems reported. Secondary level of user support at GSTT and KCL (such as login, networking and power restoration) has been excellent. There is a plan to obtain approval to measure user satisfaction.

4 Bioinformatics

The bioinformatics services activity started in September 2008 and was completed in February 2011. It is an in-house solution focusing on scientific visualisation and modelling aiming to understand research analysis and improve existing services. The use of Cloud offers two distinct advantages:

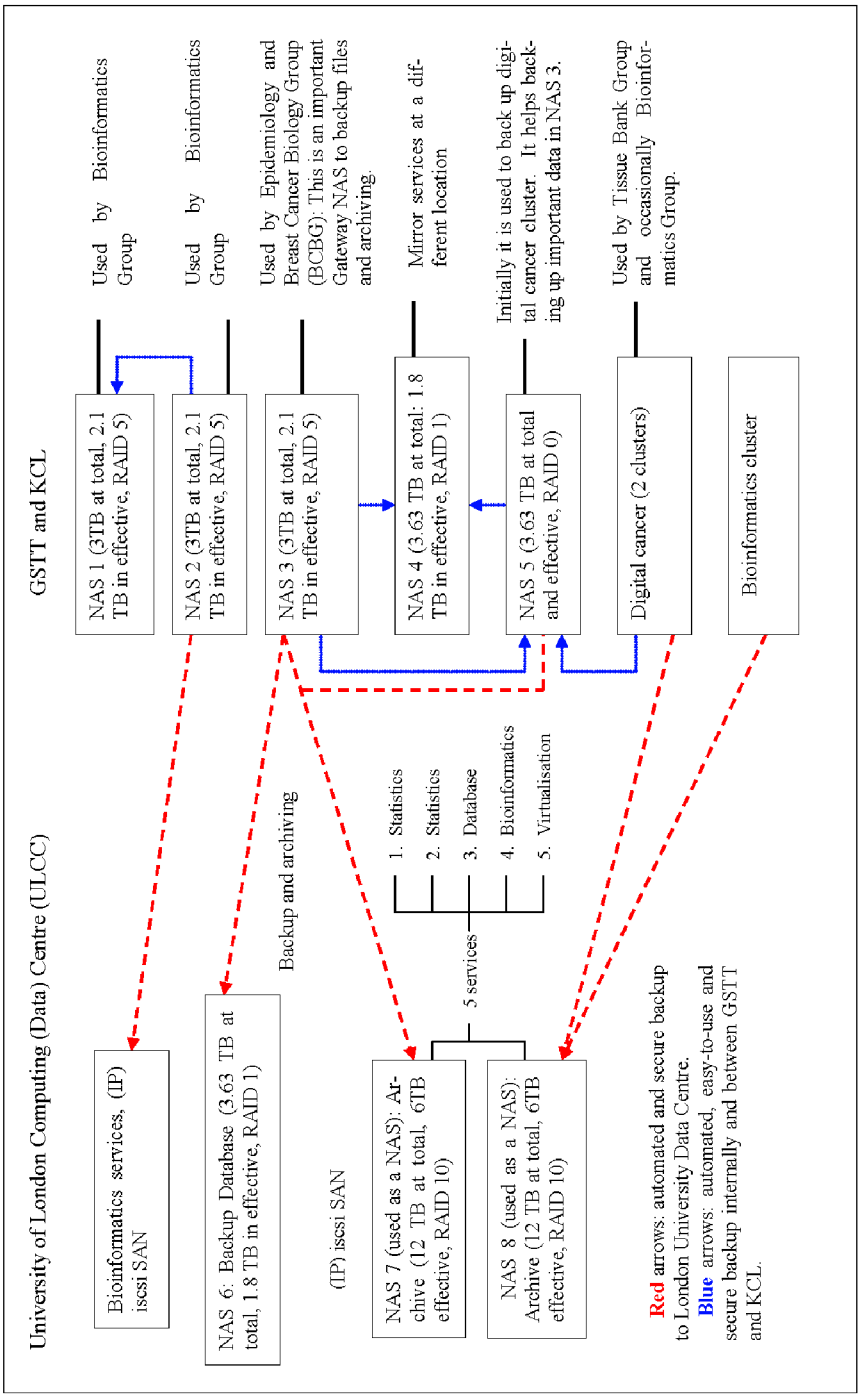

Fig. 1. Cloud Storage Deployment Architecture

(i) A PaaS for developers to simulate dynamic 3D modelling and visualisation for proteins, genes, molecules and medical imaging, where results can be instantaneous and data can be visualised, stored and shared securely.
(ii) Any complex modelling, such as growth of tumour and segmentation of brains, can be presented with the ease.

Each section is described as follows.

4.1 Tumour Modelling

Tumours develop as a result of abnormal and rapid growth of cells, and there are two types of tumours. The first type is benign tumours, which are harmless to human bodies. The second type is malignant tumours, which are malicious, should be removed and patients with them should be treated as soon as possible. Despite the fact that current technologies can take high-resolution pictures of tumours, it is extremely helpful for high performance Cloud resources to simulate the growth and formation of tumours, and this allows scientists and surgeons to diagnose possibilities of tumour growth and gain a better understanding about treatment [21]. See Figure 2 for tumour modelling.

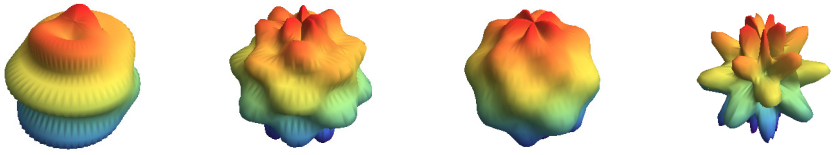

Fig. 2. Selected figures in Tumour modelling

4.2 Medical Imaging

Medical imaging is widely adopted in Hospitals and medical institutes, and new ways to improve existing medical imaging services are regularly exploited. Bioinformatics Cloud platform allows computation and visualisation, and currently brain imaging can be used for demonstration. The aim is to study segmentation of brains, which divides the brain into ten major regions. The Cloud platform has these two functions: (i) it can highlight each region for ten different segments; and (ii) it can adjust intensity of segmentation to allow basic study of brain medicine. Figure 3 below shows selected brain imaging. Segmentation is an important aspect in brain study and it has two different functionalities. Firstly, it can highlight different areas in the cerebrum, where the different light intensity can highlight which particular areas. Secondly, segmentation can show different areas in the brain, including cerebellum, temporal lobe, mid-brain and so on. This allows medical students and instructors to understand the structure of human brain with the ease, but it also provides a platform to identity the right spot of the brain in a quick and efficient manner.

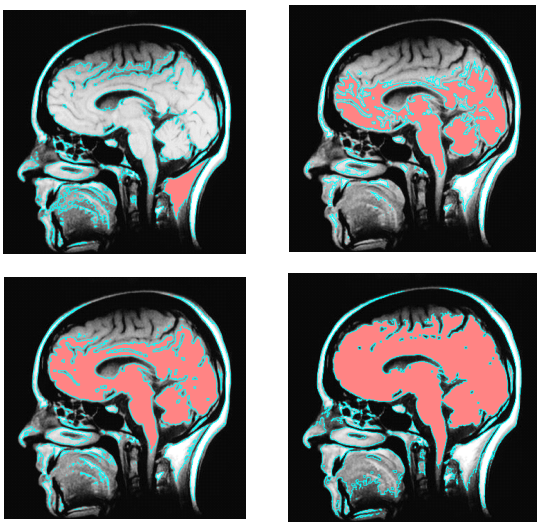

Fig. 3. Selected brain imaging

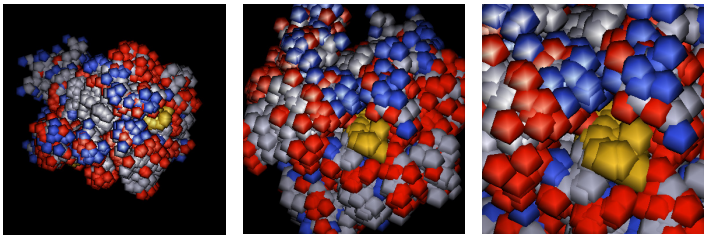

Fig. 4. Investigation of insulin molecules on Cloud

4.3 Insulin Molecules

Insulin is a hormone central to regulating carbohydrate and fat metabolism in the body, and is important for type one diabetes treatment. Insulin has a molecular structure, and the study of its structure and formation helps scientists to understand how to improve treatment. Cloud offers a platform for simulations and modelling enabling cutting-edge techniques to be used for Health Cloud for 3D Visualisation and modelling. This allows researchers to identify the areas in the molecule that they plan to study, and it allows 360 degrees rotation and zooming function, so that one particular area in the molecule can be magnified for different studies. Figure 4 shows the insulin molecule in original size and in zooms.

4.4 Simulations for Medical Training

3D simulations on Cloud are very useful for medical education and workshop, since explanations can be made easier and participants can understand better with the aid of visualisation. 3D simulations such as DNA modelling, Poyllotaxis Spirals and cleavage of embryos have been used for training, and have positive feedback and support.

5 Trials for Cloud Storage

The design and implementation of a robust Storage Area Network (SAN) requires integrations of different technologies. Only minimal modelling and simulations are needed, since the focus is on building up a service from the very beginning. Experiments provide a suitable research method, since they can identify issues such as performance, technical capabilities (such as recovery), and whether integration of technologies can deliver services. User and executive requirements are important factors for what type of experiments to be performed and measured. Thousands of files (data and records) are used for performance tests and the time to complete the same amount of jobs is recorded. Venue of test is between two sites: ULCC and GSTT/KCL and execution time is used as the benchmark. There are three data services and each service is used to perform experiments as follows:

- Backup Automation
- Data recovery
- Data migration

5.1 Backup Automation

Cloud Storage uses a number of enterprise solutions such as Iomega/EMC, Lacie, Western Digital and HP to deliver fast and reliable services including automation. The experiment performs automated backup of between 1,000 and 10,000 files, which are available in the existing system for user support. Each set of experiments is performed three times with the average time obtained. Results are shown in Figure 5.

5.2 Data Recovery

Data recovery is another important service to recover lost data due to accidents or emergency services. In the previous experience, it took two weeks to recover 5 TB of data as it required different skills and systems to retrieve data and restore good quality data back to Cloud services. Data archived as Virtual Machines or Virtual Storage speeds up recovery process. In addition, there are mirror servers so, even if a server is completely broken, data can be recovered to resume services. See Figure 6 for their execution time.

5.3 Data Migration of Single Large Files

Data migration is common amongst Clouds and is also relevant to data intensive research. When there are more organisations going for private cloud deployment, data migration between Clouds is common and may influence the service delivery [2,6,7,22]. But there is no investigation the impact of moving single large files between private clouds. Hence, the objective here is to identify the execution time for moving single large file. Each file is between 100 GB and 1 TB. Figure 7 shows the results.

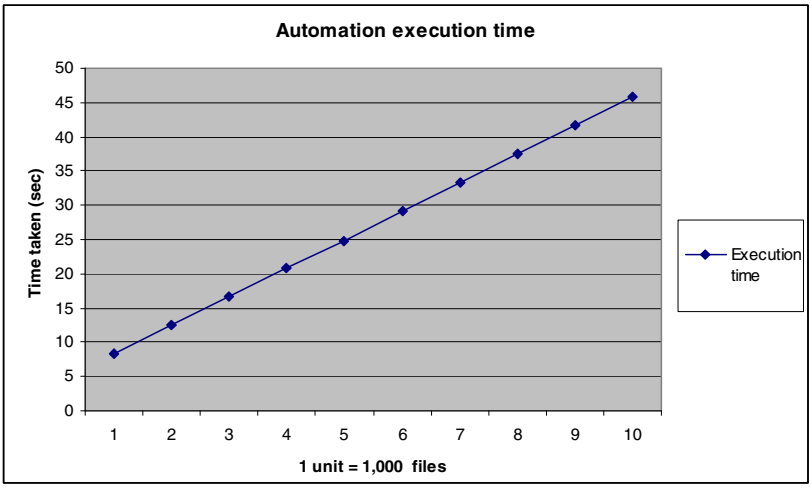

Fig. 5. Automation execution time for Cloud Storage

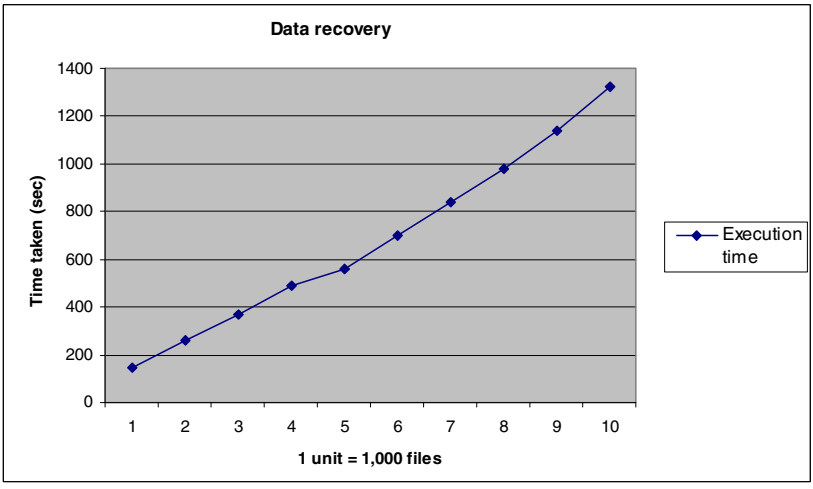

Fig. 6. Data Recovery

5.4 The Percentage of Failure Rates

The percentage of failure rates in Cloud Storage operations is important as each failure in service will result in loss of time, profit and resources. This part of experiment is to calculate the percentage of failures, where services in Section 5.1 and 5.3 are running real-time and record the number of successful and failed operations. Failed operations happen in the Cloud environments. Monitoring the failure rate is important as failures contribute to the development of risks. To reduce the impacts

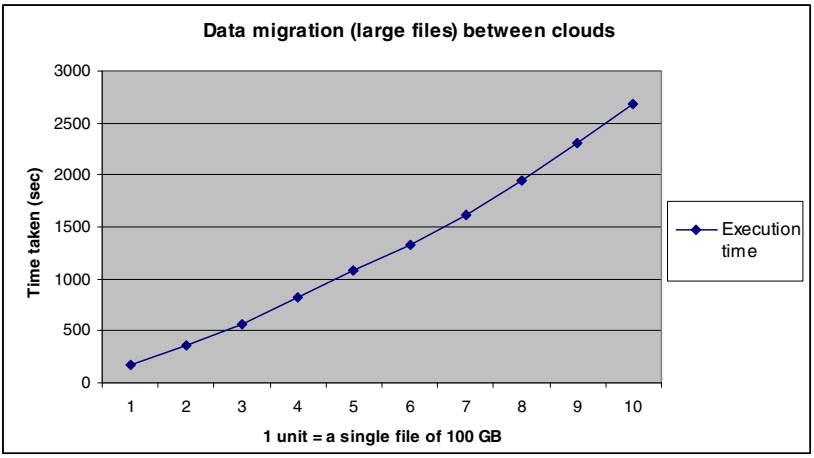

Fig. 7. Data migration of large single files between clouds

from risk (as a result of Cloud adoption), Chang et al [12,13,16,20] demonstrate that controlled risk in Cloud adoption can be monitored and presented in the form of risk-free rate, or risk-occurring rate if the focus is on the measuring the extent of failure rates. There are hundreds of successful operations versus and a number of failed operations.

5.4.1 Failure Rate in Backup Automation

Backup automation is relatively reliable and out of hundreds of thousands of operations, the failure rate is below 2%. The reason is that backup automation has been available for a significant number of years with the result that it is a mature technology

5.4.2 Failure Rate in Data Recovery

Data recovery for large-scale data in Cloud is important and the failure rate is shown in Figure 8 based on the number of successful and failed operations since 2009. The interesting result is when there is a low amount of data, the percentage of failure is low. When the amount of recovered data increases, the execution time is approximately proportional to the amount of data but the failure rate increases more quickly and the graph looks close to an exponential curve.

5.4.3 Failure Rate in Data Migration

Data migration of large files in Cloud is common and important as Storage is designed for terabytes and petabytes. The failure rate is shown in Figure 9 based on the number of successful and failed operations since 2009. Similar to Figure 8, the curve is close to an exponential one, which means when the volume of the migrated file increases, the failure rate increases significantly.

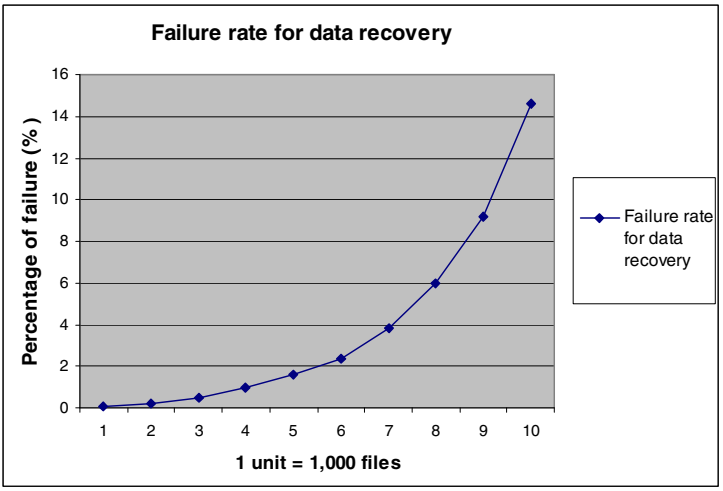

Fig. 8. Failure rate of data recovery

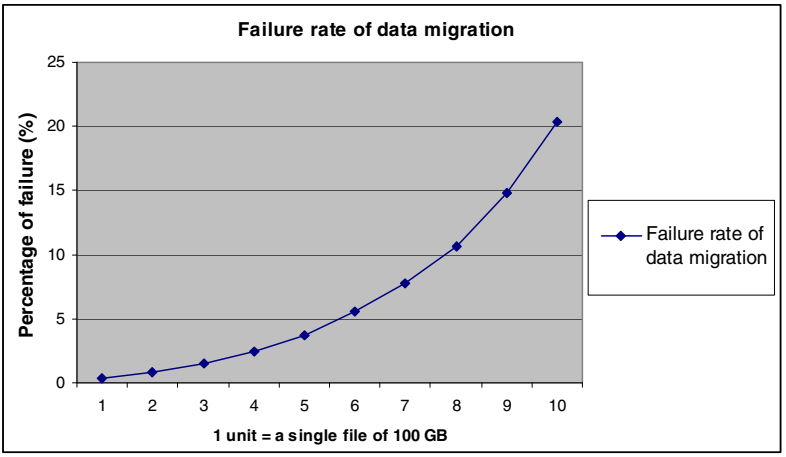

Fig. 9. Failure rate of data migration

5.5 Summary of All Experiments

Service and backup automation for Cloud storage takes the least execution time and there are several services to speed up the process of automation. Execution time is between 8 and 46 seconds backup 1,000 to 10,000 files to automatically. The second experiment is data recovery, where data archived as Virtual Machines or Virtual Storage in a well-managed platform can speed up recovery process. Data recovery takes between 135 seconds to 1,312 seconds to recover 1,000 to 10,000 of files. The third experiment focuses on data migration of large single files, which are important for data intensive research. Data migration takes between 174 seconds to 2,686

seconds to move a single file of 100 GB to 1 TB. Although Figures 3 and 4 still show a linear graph, more execution time is required to recover data and move a large single file and the percentage of unsuccessful data recovery and migration is likely to increase.

The results strongly suggest that it is quicker to move data around Clouds in many smaller files. Our results also confirm that automation in Cloud is better established than data recovery and data migration of single large files, and these two are perhaps challenges that data-intensive research need to overcome. Failure rate for these three major operations are demonstrated. Backup automation is the most reliable and stays below 2% all the times. Figure 6 is similar to Figure 5 and shows that failure rate of data migration; when the volume of the migrated file increases, so does the failure rate.

6 Discussions

There are several topics for discussions presented as follows.

6.1 Challenges for Data Intensive Research in Cloud

Cloud Storage can offer services up to petabytes of storage and beyond. The results in Section 4 confirm that large-scaled data recovery and data migration in Cloud need to improve in its technical capabilities. This is reflected in the percentage of failure rate and how the failure rate apparently increases exponentially to 14.6% as data recovery volume increases to 10,000 files. Similarly, an exponential increase is experienced to 20.4% when data migration disk increases to 1 TB per file. Our results demonstrate data recovery and data migration issues for thousands of files have to be resolved and improved prior dealing with challenges in petabytes.

6.2 User Feedback on Cloud Storage

Currently Cloud Storage has provided users the following benefits:

- Cost reduction: The service is automated and saves costs in hiring and deploying staff and deployment of a larger and more expensive project that works the same. There is no need to hire a team to look after maintenance and daily services.
- Time-saving: Cloud Storage simplifies the complex backup process and saves time in performing backups. Users find that they need not spend significant time for back up.
- User friendliness: Cloud Storage offers easy to use features and users without prior knowledge can find it simple to use.

Healthcare community has a Data Protection Policy and not all types of services are able to release data. Services that do not use patients' data or confidential information are likely to be presented.

6.3 Plug and Play Features in Cloud Storage for Data Intensive Research

There are papers explaining the importance and relevance of data intensive research, and why it is essential for Cloud development and services [22,26]. This Cloud Storage allows plugs and play, which means adding additional hard disks to existing NAS, or new NAS, can still provide services in place. This has been tested in 2010 where disk volume of NAS 7 and 8 were increased from 20 TB to 44 TB without interruptions of services. This Cloud Storage was also tested to store and protect data of up to 100 TB on another occasion. This allows any addition of hard disks and applications within 100 TB limit to provide user support and services.

Cloud Storage has been in used daily by medical researchers, and there are a few local administrators supporting a minimum level of services. The focus for this service is no longer in technical implementation but rather user satisfaction.

6.4 Relative Performance

Buyya et al [5,6,7] describe technical performance in detail. Often results are very technical and most organisations considering or implementing Clouds find those results difficult to follow [10,11,13]. Relative performance is an easier term to compare performance with, and is defined as the improvement in performance between an old service (before) and a new service (after). Latch et al. [24] also use relative performance to present their Bayesian clustering software where the key performance indicators are presented in terms of percentages of improvement. Although Latch et al. [24] still use statistical approach where some data have little impact or relevance to organisational adoption, the benefit of using relative performance approach is to bring down level of complexity and allows stake holders to understand the percentage of improvement.

A hybrid case study is relevant for organisational Cloud adoption, since data needs to be checked prior computational analysis and often this needs supporting interviews and surveys. From interviewing members of management, their views can be summed up as follows:

- They support the use of relative performance, as most of the executives are not from IT backgrounds.
- The use of key performance indicators in relative performance makes it easy for the executives to understand and follow the extents of improvement.

6.5 The Proposal for "Healthcare Platform as a Service" (Hpaas) for Research and Education

Cloud Computing offers contributions to research and development, as complex simulations can be computed and modelled with the on-demand capabilities, elasticity and scalability that Cloud can provide. Genes, molecules and medical imaging can be modelled at high speed and results can be computed and viewed in real-time. This is due to the establishment of PaaS to minimise the execution time so that 3D simulation can be running right after the code development on Cloud.

Bioinformatics services also compare the performance improvements before and after introducing Cloud as an important ROI measurement. Chang et al. [12] demonstrate that 1.2% - 7.2% time reduction for code development is achieved. Their objective is clearly met and project delivery is straightforward with progressive improvements. Different Health Cloud projects in Infrastructure, Bioinformatics, Statistics, HPC, Data Services and Security have worked together in an integrated environment to establish Health Platform as a Service (HPaas), which brings the following benefits:

- Different activities in private cloud can work together.
- The expertise in each area can be consolidated within the HPaaS.
- The outcome of one service can be the input to another.

Efficiency has improved as the Cloud saves time and resources to repeat the same processes, which can be automated. This is important in case the systems and/or services break that automated virtualised environments can quickly provision to the original setting. 3D Bioinformatics enhances the level of research and simulations can help surgeons and medical staff to make the right decisions. Chang et al. [14,17,20] also demonstrate Business Integration as a Service (BIaaS) that can further improve the process and integration of different activities in HPaaS.

7 Conclusions

This paper illustrates PaaS Portability in the form of Healthcare Cloud Storage, which is designed, deployed and serviced to GSTT and KCL under the recommendation of CCAF to ensure good Cloud design, deployment and services. Service Portability has been designed, implemented and serviced at participating organisations to provide added values such as efficiency improvement and time reduction in code development and execution time. User Groups for the system are divided into Bioinformatics Group, Databank and Cancer Epidemiology Group, BCBG Group, Tissue Bank and Senior Clinicians. The CCAF was useful and helped the Health Community to achieve good private cloud design, deployment and services while following user requirements and challenges, and executives' feedback closely.

Healthcare Cloud Storage implements a data service as an easy-to-use, automated and collaborative platform which some users use every day. It is distributed between two physical locations: University of London Data Centre and GSTT/KCL and is designed and built to align with group and research requirements. It uses a private-cloud SAN architecture made up from different NAS services.

The Deployment Architecture shows the connections between different NAS services and how they are related. These services include Bioinformatics (multiple services), joint Epidemiology and BCBG service, mirror services, two archiving services, digital cancer services and multiple backup services. Automated and secure backups take place between the two physical locations.

The first lesson from this activity is that recommendations from CCAF assist with achieving good Cloud Design. A further lesson is that using experiments when designing and implementing a Cloud-based Storage Area Network (SAN) is helpful

and execution time can be used as the benchmark to determine their success. Experiments were performed in three areas: automation, data recovery and data migration.

- Automation in Cloud storage has enabled several services to speed up the process of automation. Execution time is between 8 and 46 seconds to automate backup 1,000 to 10,000 files.
- Data recovery in a well-managed platform can speed up recovery process and takes between 135 seconds to 1,312 seconds to recover 1,000 to 10,000 of files. Data migration of large single files is important for data intensive research. Data migration takes between 174 seconds to 2,686 seconds to move a single file of 100 GB to 1 TB.
- Our results also confirm that backup automation in Cloud is more mature than data recovery or data migration of single large files, and these two represent challenges that data-intensive research needs to overcome. Relative performance is between Cloud Storage and traditional storage have been presented.

Percentage failure rate is calculated for backup automation, data recovery and data migration. Backup automation failure rate stays below 2% but the failure rate increases rapidly to 14.6% for data recovery as the volume increases to 10,000 files. Similarly, a rapid increase to as much as 20% is experienced in data migration as data migration file size increases towards 1 TB. These results suggest that issues and challenges remain within data recovery and migration which will need to be resolved before systems progress to handling petabytes of storage.

Healthcare platform (HPaaS) enables different activities to work together, so that expertise in one area can be consolidated. The use of 3D simulations allows developers to compute results in real-time and data can be stored, visualised and shared securely. 3D simulations of tumour, medical imaging and insulin have also helped to improve the quality of research analysis, as well as providing better understanding in the structure and formation of these analyses. All complex life science modelling can be presented with ease, so that it not only can promote greater awareness of health and disease issue, but also improves the quality of current research and development.

References

1. Abu-Libdeh, H., Princehouse, L., Weatherspoon, H.: RACS: A Case for Cloud Storage Diversity. In: SoCC 2010 Proceedings of the 1st ACM Symposium on Cloud Computing, Indianapolis, Indiana, June 10-11 (2010)
2. Armbrust, M., Fox, A., Griffith, R., Jseph, A.D., Katz, R.H., Kownwinski, A., Lee, G., Patterson, D., Rabkin, A., Stoica, I., Zaharia, M.: Above the Clouds: A Berkeley View of Cloud computing. Technical Report, No. UCB/EECS-2009-28, UC Berkeley (February 2009)

3. Beaty, K., Kochut, A., Shaikh, H.: Desktop to Cloud Transformation Planning. In: 2009 IEEE International Symposium on Parallel and Distributed Processing, Rome, Italy, May 23-May 29 (2009)
4. Bryant, R.E.: Data-Intensive Supercomputing: The Case for DISC, Technical paper, Carnegie Mellon University (October 2007)
5. Buyya, R., Yeo, C.S., Venugopal, S., Broberg, J., Brandic, I.: Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility. Journal of Future Generation Computer Systems 25(6), 559–616 (2009)
6. Buyya, R., Ranjan, R., Calheiros, R.N.: InterCloud: Utility-Oriented Federation of Cloud Computing Environments for Scaling of Application Services. In: Hsu, C.-H., Yang, L.T., Park, J.H., Yeo, S.-S. (eds.) ICA3PP 2010, Part I. LNCS, vol. 6081, pp. 13–31. Springer, Heidelberg (2010)
7. Buyya, R., Beloglazov1, A., Abawajy, J.: Energy-Efficient Management of Data Center Resources for Cloud Computing: A Vision, Architectural Elements, and Open Challenges. In: Buyya, et al. (eds.) PDPTA 2010 - The International Conference on Parallel and Distributed Processing Techniques and Applications, Las Vegas, USA, July 12-15 (2010b)
8. Chang, V.: Cloud Storage Framework – An Integrated Technical Approach and Prototype for Breast Cancer., Poster Paper and Technical Paper, UK All Hands Meeting (December 2009)
9. Chang, V., Bacigalupo, D., Wills, G., De Roure, D.: A Categorisation of Cloud Computing Business Models. In: Chang, et al. (eds.) The 10th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing, CCGrid 2010, Melbourne, Australia, May 17-20, pp. 509–512 (2010a)
10. Chang, V., Wills, G., De Roure, D.: A Review of Cloud Business Models and Sustainability. In: Chang, et al. (eds.) The Third International Conference on Cloud Computing, IEEE Cloud 2010, Miami, Florida, USA, July 5-10 (2010b)
11. Chang, V., Li, C.S., De Roure, D., Wills, G., Walters, R., Chee, C.: The Finan-cial Clouds Review. International Journal of Cloud Applications and Computing 1(2), 41–63 (2011a) ISSN 2156-1834, eISSN 2156-1826
12. Chang, V., De Roure, D., Wills, G., Walters, R., Barry, T.: Organisational Sus-tainability Modelling for Return on Investment: Case Studies presented by a National Health Service (NHS) Trust UK. Journal of Computing and Information Technology 19(3) (2011b) (in press); ISSN Print ISSN 1330-1136 | Online ISSN 1846-3908
13. Chang, V., De Roure, D., Wills, G., Walters, R.: Case Studies and Organisational Sustainability Modelling presented by Cloud Computing Business Framework. International Journal of Web Services Research (2011c) (in press) ISSN 1545-7362
14. Chang, V., Wills, G., Walters, R.: Towards Business Integration as a Service 2.0 (BIaaS 2.0). In: Chang, et al. (eds.) IEEE International Conference on e-Business Engineering, The 3rd International Workshop on Cloud Services - Platform Accelerating e-Business, Beijing, China, October 19-21 (2011d)
15. Chang, V., Wills, G., Walters, R.: The positive impacts offered by Healthcare Cloud and 3D Bioinformatics. In: Chang, et al. (eds.) 10th e-Science All Hands Meeting 2011, York, September 26-29 (2011e)
16. Chang, V., Wills, G., Walters, R., Currie, W.: Towards a structured Cloud ROI: The University of Southampton cost-saving and user satisfaction case studies. In: Chang, et al. (eds.) Sustainable Green Computing: Practices, Methodologies and Technologies (2012a)
17. Chang, V., Walters, R., Wills, G.: Business Integration as a Service. International Journal of Cloud Applications and Computing 2(1) (2012) ISSN 2156-1834, eISSN 2156-1826

18. Chang, V., Walters, R.J., Wills, G.: Cloud Storage in a private cloud deployment: Lessons for Data Intensive research (Best student paper). In: Chang, et al. (eds.) The Second International Conference on Cloud Computing and Service Sciences (CLOSER 2012), Porto, Portugal (2012c)
19. Chang, V., Wills, G.: A University of Greenwich Case Study of Cloud Computing – Education as a Service. In: E-Logistics and E-Supply Chain Management: Applications for Evolving Business. IGI Global (2013)
20. Chang, V., Walters, R.J., Wills, G.: The development that leads to the Cloud Computing Business Framework. International Journal of Information Management (February 2013)
21. Grigoriadis, A., Chang, V., Schuitevoerder, M., Gillet, C., Tutt, A., Holmberg, L.: Cancer Cloud Computing - Towards an Integrated Technology Platform for Breast Cancer Research., Internal NHS Technical Paper (July 2009)
22. Hey, A.J.G.: The fourth paradigm: data-intensive scientific discovery. Microsoft Publication (2009) ISBN-10: 0982544200
23. Kagermann, H., Österle, H., Jordan, J.M.: IT-Driven Business Models: Global Case Studies in Transformation. John Wiley & Sons (2011)
24. Latch, E.K., Dharmarajan, G., Glaubitz, J.C., Rhodes. Jr., O.E.: Relative performance of Bayesian clustering software for inferring population substructure and individual assignment at low levels of population differentiation. Conservation Genetics 7, 295–302 (2006), doi:10.1007/s10592-005-9098-1
25. Moore, R.W., Baru, C., Marciano, R., Rajasekar, A., Wan, M.: Data-Intensive Computing. In: The Grid: Blueprint for a New Computing Infrastructure, ch. 5 (1999) ISBN 1558609334
26. Moretti, C., Bulosan, J., Thain, D., Flynn, P.J.: All-Pairs: An Abstraction for Data-Intensive Cloud Computing. In: IEEE International Symposium on Parallel and Distributed Processing, IPDPS 2008, Miami, USA, April 14-18 (2008)
27. Vo, H.T., Chen, C., Ooi, B.C.: Towards Elastic Transactional Cloud Storage with Range Query Support. Proceedings of the VLDB Endowment 3(1-2) (September 2010)