Cloud Storage in a private cloud deployment: Lessons for Data Intensive research

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

[1] School of Electronics and Computer Science, University of Southampton, Southampton SO 17 1BJ, UK.
{vic1e09, rjw1, gbw}@ecs.soton.ac.uk
[2] Business School, University of Greenwich, London SE10 9LS, UK
{cv24}@gre.ac.uk

Abstract. This paper demonstrates portability for a private cloud deployment, which has a detailed case study about Cloud Storage service developed as part of the Cloud Computing Business Framework (CCBF). 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. Data recovery result confirms that execution time is in proportion to quantity of recovered data, but the failure rate increases in an exponential manner. Data migration result confirms execution time is in proportion to disk volume of migrated data, but the failure rate increases in an exponential manner. Issues in data recovery and data migration must be resolved prior dealing with petabytes of data. Our Cloud Storage offers cost reduction, time-saving and user friendliness supported by users and is highly relevant to similar portability of private cloud.

1 Introduction

Communications between different types of clouds from different vendors are often difficult to implement. Often work-arounds require writing additional layers of APIs, or an interface or portal to allow communications. This brings interesting research question such as portability, as portability of some applications from desktop to cloud is challenging (Beaty et al., 2009; Armbrust et al., 2009). Portability refers to moving enterprise applications and services to Clouds from all types, and not just files or VM over clouds. Beaty et al. (2009) and Chang et al. (2011 a) identify portability as an organisational challenge for Cloud adoption and explain their rationale and demonstration in their paper.

This paper presents portability in Healthcare, where Cloud Computing Business Framework (CCBF) is a framework that has been involved from service strategy to design, development, test and user support stages. There is a Cloud Storage project focusing on design, implementation, and user support and is the focus for this paper. Details for Cloud Storage include functionalities, technical implementation, architecture and user support. Experiments for data services (backup automation, data recovery and data migration) are performed and results can help us to meet issues and challenges of data-intensive research.

1.1 Enterprise Portability

Enterprise portability (portability in short) involves moving applications and services from desktops to clouds and between different clouds, including IaaS, PaaS and SaaS implementations. This is domain specific as there are different requirements for portability in each domain. This paper describes examples in Healthcare. These Cloud projects have been successfully delivered and provide a high level of user satisfaction.

CCBF aims to help organisations to achieve good Cloud design, implementation and services (Chang et al., 2011 b; 2011 c; 2011 d; 2011 e; 2011 f). Portability plays an influential role in two aspects for Healthcare:

- Migrating existing infrastructure, platform and applications to Cloud environments.
- Developing a new platform and/or application to allow new development of Cloud services.

Cloud storage is an in-house private cloud initiative and the initial focus is to build a working IaaS infrastructure to allow storage of medical databases, images and analysis in a secure and collaborative environment. After spending a period with smooth delivery and user support, the focus becomes upgrading from IaaS to PaaS, which allows better benefits such as better efficiency and better management of resources. 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 discusses its performance results, Section 5 presents topics of discussion and Section 6 sums up Conclusion and future work.

2. Cloud Storage

Cloud Storage is a crucial project funded and supported by NHS UK, where Guy's and St Thomas NHS Trust (GSTT) and King's College London (KCL) have worked together to deliver a service, with the initial plan for proof of concepts and to see whether Cloud Storage can be useful. Cloud Storage development began in September 2008 and completed in May 2010 to serve cancer researchers. CCBF is instrumental and influential in the way Cloud Storage has been developed with the following reasons:

- Cloud Storage is a PaaS, and needs careful planning and a thorough implementation. This requires using an integrated adoption of multiple vendors' solutions.
- Cloud Storage is an area to experience rapid growth in user requirements and disk space consumption. Therefore, it must be easy to use, and able to cope with an increasing demand.
- Cloud Storage is a new concept and implementation for Health domain, as in the past, private and in-house storage is used. Maintenance of data protection and security is a challenge. Recommendation, strategy and support from CCBF can offer useful guidelines and good services.

The Healthcare Cloud Storage is used in 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 thousands of data about patients (medical records) and tumours (detailed descriptions and images, and its relations to the patients). Data growth is rapid and 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 and robust. There are also uses of third party applications to allow researchers to be able to access, view and edit any tumour images from trusted places. Security has been enforced in terms of data encryption, SSL and firewall. This project is considered as a group of Private Clouds, and is not yet to join all different Private Clouds (distributed in different areas) together.

Cloud Storage adopts Experiments as the main method to design and implement a robust Storage Area Network (SAN). This is based on integrations of different technologies where experiments are required. In relation to Cloud portability, better performance in Cloud Storage than outdated storage service is regarded as a benchmark and measurement for success by executives. A hybrid case study is used, as it requires both quantitative and qualitative information not covered by experiments. Occasionally checking with users and executives about their requirements and services they prefer, take place in the form of interviews, and thus a certain extent of qualitative method is needed. Users are very supportive in this project and some of them use it daily.

2.1 Benefits and impacts of adopting CCBF

The benefits of adopting CCBF allows their IT lead to understand requirements, technical knowledge, use cases and issues to be aware of, before and during the project development. The Private Cloud Storage project 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 CCBF, and carry out plans for building and validating the framework.
- Stage 3: Propose and implementing service oriented architecture for Cloud Storage based on CCBF. Offer services for users and research groups.

- Stage 4 (Current stage): Continue for service improvements and provide integrations with other services or new requirements.

The features in the Healthcare Cloud Storage include the followings:

- Automation of backup services
- Easy backup and archiving
- Snapshots and mirroring
- Replication
- Recovery
- Data Migration
- Test-bed / test environments
- Heterogeneous network and OS support
- Proof of concepts
- Some services are offering user support

It offers a wide range of self- and automated services across secure networks. There are also options for retrieving and viewing data through the Intranet and Internet but are only accessible if a secure VPN and Secure Sockets Layer (SSL) certificates are used. The CCBF positively influences the way the backup and storage are designed and deployed. This project involves from the following:

- Building infrastructures in IaaS.
- Implementing, upgrading and testing systems and resources to PaaS level.
- Resolving existing problems and making improvements Cloud Storage services.

This needs the state-of-the-art design and implementation that the CCBF can offer. There are two different focuses. Firstly, it must be easy to use and support several research groups synchronously and asynchronously. The Cloud Storage must be able to cope with frequent changes, updates and user activities. Secondly, the platform must be highly robust and stable, and allow data to be kept safe, secure and active for a long period of time, in other words, ten years and above. This allows data archiving, mirroring and recovery. 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.

The Architecture design is decided to build and support 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 supports active user activities, and provides great usability and accessibility while maintaining a high level of system architecture, programming and maintenance work. The NAS supports individual backups with manual and automated options. One option is similar to the Dropbox pattern of backup, and users can copy their files onto their allocated disk space, without worrying about complexity because backup is easy to use and user-friendly. The manual backup option allows users to backup their resources onto a selected destination, and offer both compressed and uncompressed versions of backup. There are options to choose data encryption to enforce security.

2.2 A Storage Area Network made up of different clusters of Network Attached Storage (NAS)

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, and RAID 10 is a good choice since it can boost the performance like RAID 0, and it 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.

The CCBF approach 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. Some applications mainly based on PHP, MySQL and Apache are 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. 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 (1999) and Bryant (2007) 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 to Cloud Storage as the design and deployment must provide resilient, swift and effective services. Vo, Chen and Ooi (2010) present their Cloud Storage and experiment their read, write and transaction throughput. They demonstrate their solution for data migration but there is a lack of consideration in data recovery which is important in the event of possible data loss. Abu-Libdeh, Princehouse and Weatherspoon (2010) demonstrate their Cloud Storage case study which presents how “Failure Recovery” that 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 4.
Infrastructure	Collaborator and	University of	Some services need a more	Amount of work is reduced for

(networking and hosting solution)	in-house	London Data Centre	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.	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.

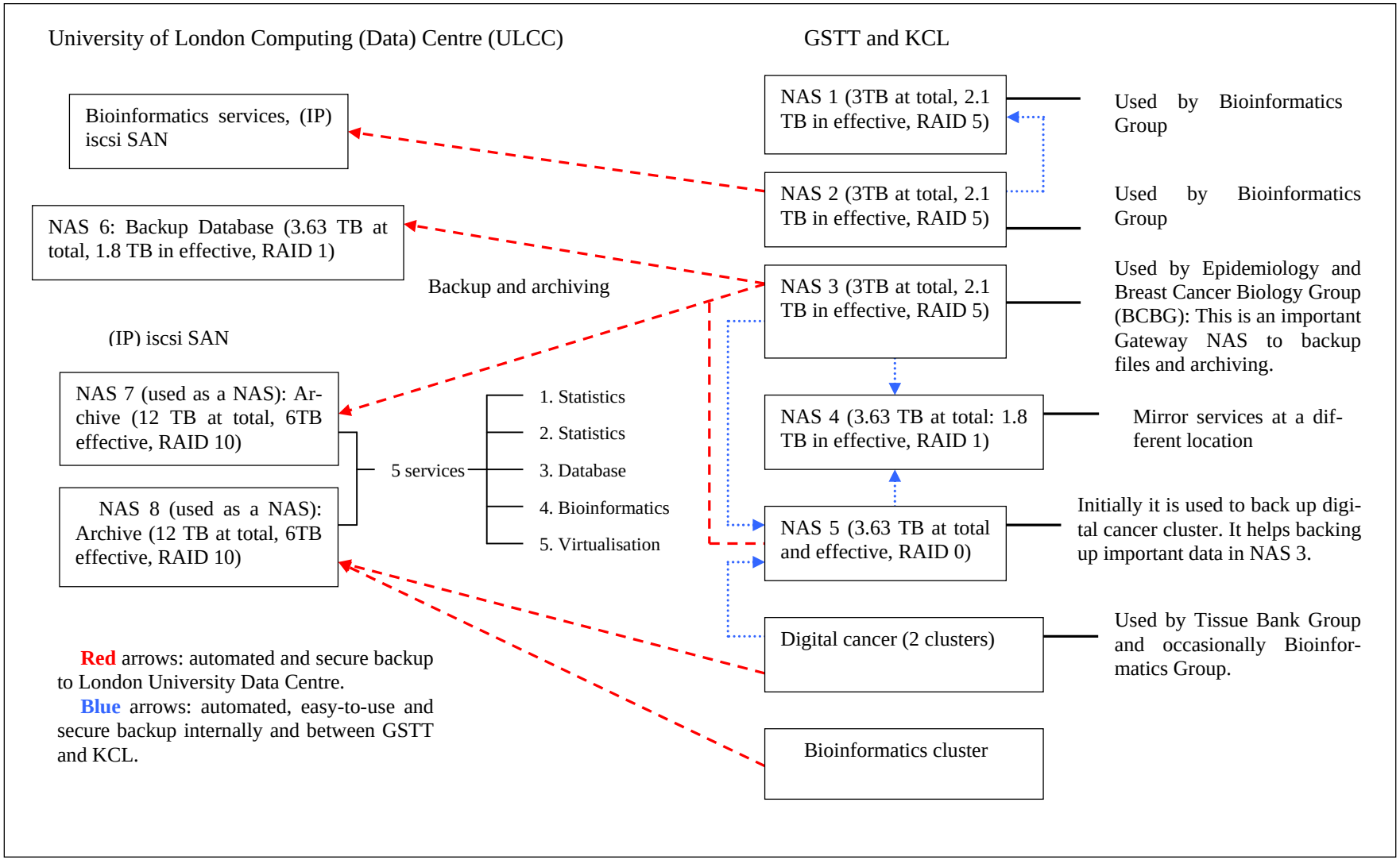
5.2 Deployment Architecture

There are two sites for hosting data, one is jointly at GSTT and KCL premises distributed in dedicated server rooms and the other is hosted 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, where 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 to do this is because disaster recovery took place in 2010 and that 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 are excellent and most data can be backed up in a short and acceptable time frame for not more than 1 hour to back up thousands of 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 on Database to store confidential data and the fourth is on bioinformatics to help bioinformatics research, and 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.

Figure 1: Cloud Storage Deployment Architecture



3.3 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. Experiments for Cloud Storage

Design and implement 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 are the suitable research method, since it 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

4.1 Backup Automation

Cloud Storage uses a number of enterprise solutions such as Iomega/EMC, Lacie, Western Digital and HP to deliver a fast and reliable automation services. The experiment performs backup automation between 1,000 and 10,000 files, which are available in the existing system for user support. The benchmark is dependent on execution time. Each set of experiments is performed three times with the average time obtained. Time taken was recorded to present all results as shown in Figure 2.

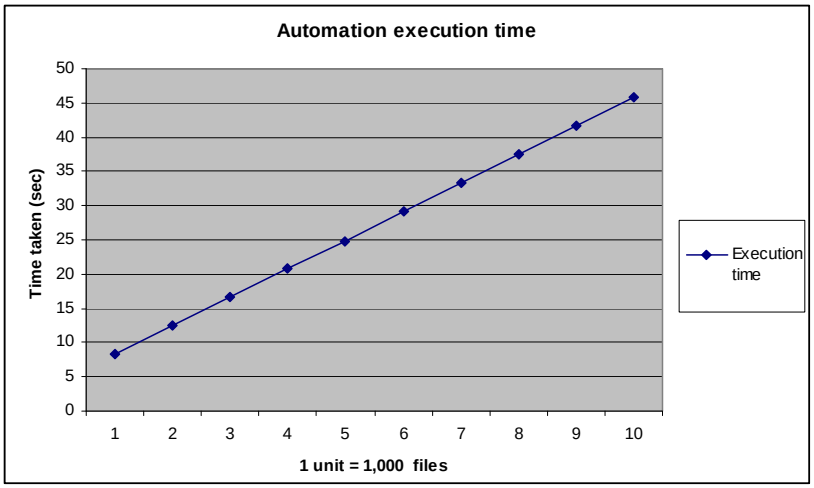


Figure 2: Automation execution time for Cloud Storage

4.2 Data Recovery

Data recovery is an 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 for disaster recovery as it requires 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, and even if a server is completely broken, data can be recovered to resume services so that recovery does not take days and weeks. See Figure 3 for their execution time.

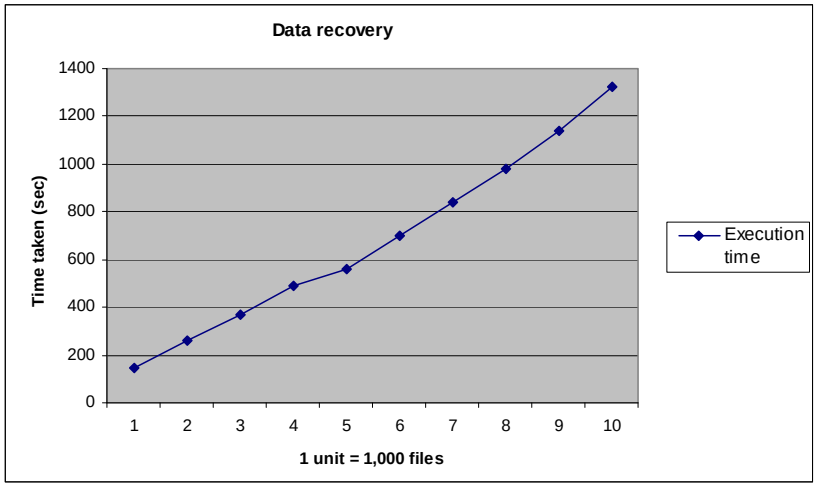


Figure 3: Data Recovery

4.3. 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 way service delivery (Ambrust et al., 2009; Hey, 2009; Buyya et al; 2010 a; 2010 b). But there is no investigation the impact of moving single large files between private clouds. Hence, the objective is to identify the execution time for moving single large files and each file is between 100 GB and 1 TB. Figure 4 shows the results.

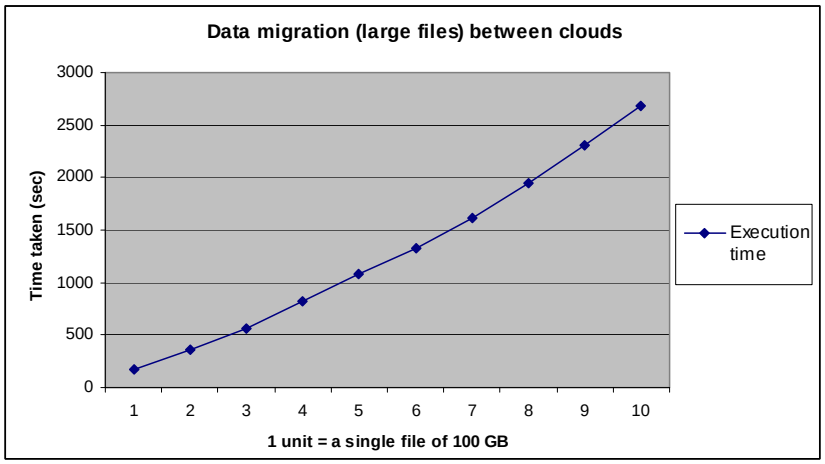


Figure 4: Data migration of large single files between clouds

4.4 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 4.1 and 4.3 are running real-time and record down the number of successful and failed operations. There are hundreds of successful operations versus and a number of failed operations.

4.4.1 Failure rate in backup automation

Backup automation is relatively reliable and out of hundreds and thousands of operations, the failure rate is below 2%. The main reason is backup automation has been in the Storage for a significant number of years and this area is more established.

4.4.2 Failure rate in data recovery

Data recovery for large-scale data in Cloud is important and the failure rate is shown in Figure 5 based on the amount 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 failure rate increases where the graph looks close to an exponential curve. This may mean the more recovered data in the Cloud, even though the execution time is in proportion to quantity, the failure rate increases in an exponential manner.

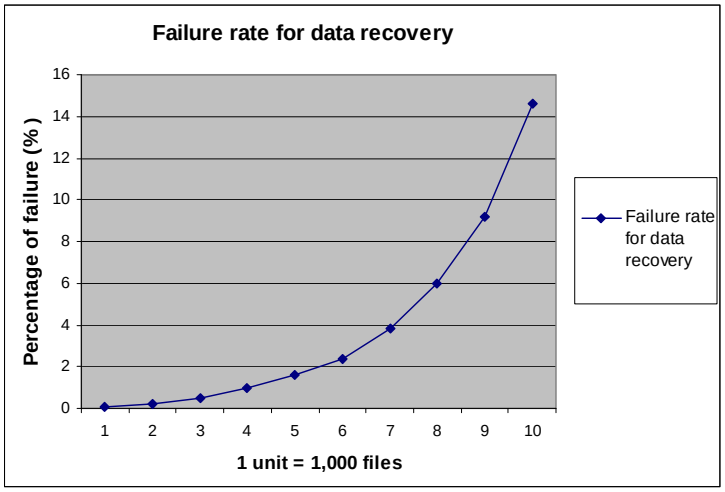


Figure 5: Failure rate of data recovery

4.4.3 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 6 based on the amount of successful and failed operations since 2009. Similar to Figure 5, the curve is close to an exponential one, which means when the volume of the migrated file increases, the failure rate increases.

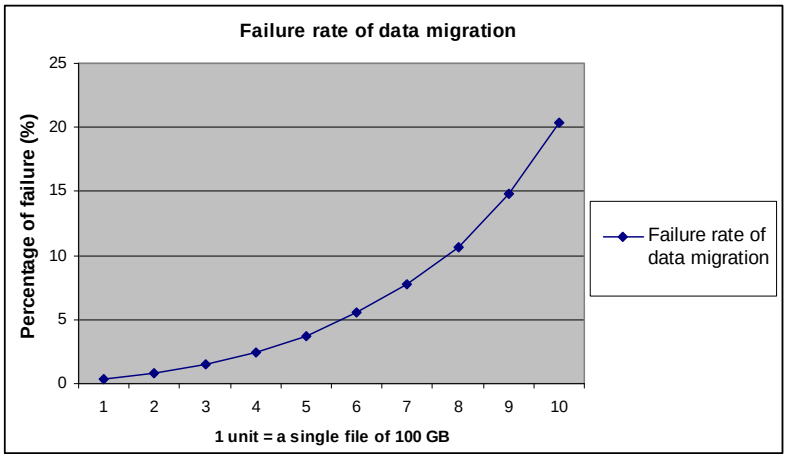


Figure 6: Failure rate of data migration

4.5 Summary of all experiments

This section is to discuss results in the previous three 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 to automate backup 1,000 to 10,000 files. 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 Figure 3 and 4 still show a linear graph,

more execution time is required to recover data and move a large single file. The percentage of unsuccessful data recovery and migration is likely to increase.

The results strongly support that it is quicker to move more data with smaller size than to move less data with larger file size in Clouds. Our results also confirm that automation in Cloud is more 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 5 shows the failure rate of data recovery and when the amount of recovered data increases, the failure rate increases where the graph looks close to an exponential curve. Figure 6 is similar to Figure 5 and shows that failure rate of data migration, which means when the volume of the migrated file increases, the failure rate increases.

5. Discussions

There are several topics for discussions presented as follows.

5.1 Challenges for data intensive research in Cloud

Cloud Storage can offer services up to petabytes of storage. The interesting results in Section 4 confirm that large-scaled data recovery and data migration in Cloud needs to improve in its technical capabilities. This is reflected in percentage of failure rate, where failure rate increases like an exponential manner up to 14.6% when data recovery volume increases up to 10,000 files. Similarly, an exponential increase is experienced when data migration increases up to 20.4% when data migration disk increases up to 1 TB per file. Our results demonstrate data recovery and data migration for thousands of files (that each has up to 1TB) have to be resolved and improved prior dealing with challenges in petabytes. Our experiments and results are not only applicable locally but also are applicable in other environments.

5.2 User feedback on Cloud Storage

Currently Cloud Storage is only active service that has been used daily, and has provided users the following benefits.

- Cost reduction: The service is automated and saves costs in hiring additional 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 friendly: 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.

5.3 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 (Moretti et al., 2008; Hey, 2009). This Cloud Storage allows plugs and plays, 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 in services without interruptions of services. This Cloud Storage was also tested to store and protect data of up to 100 TB in another occasion. This allows any addition of hard disks and applications within 100 TB limit to provide user support and services.

5.4 Current Status for Cloud Storage

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 not longer in technical implementation but rather user satisfaction. This needs to write to the NHS UK and get their support to follow up this research.

5.5 Relative performance

There are papers describing technical performance in detail (Buyya et al 2009, 2010 a; 2010 b). Often results are very technical and most of organisations considering or implementing Clouds find those results difficult to follow (Chang 2010 d, 2011 a, 2011 c). 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). This is similar to Organisational Sustainability Modelling (OSM) where data is compared between 'before' and 'after' in the areas of technical, costs and user aspects. Latch et al. (2006) also use relative performance to present their Bayesian clustering software where the key performance indicator is presented in terms of percentages of improvement. Although Latch et al. (2006) 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 interviews or surveys to support. 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 indicator in relative performance makes it easy for the executives to understand and follow the extents of improvement.

6. Conclusion and Future Work

This paper illustrates PaaS Portability in the form of Cloud Storage, which is designed, deployed and serviced to GSTT and KCL under the recommendation of CCBF to ensure good Cloud design, deployment and services. Cloud Storage is helpful to make data service an easy-to-use, automated and collaborative platform and some users use this service daily. Design and deployment of Cloud Storage have been described, and have followed user requirements and executives' feedback closely. User Groups are divided into Bioinformatics Group, Databank and Cancer Epidemiology Group, BCBG Group, Tissue Bank and Senior Clinicians. The best approach is to design and implement a Cloud-based Storage Area Network (SAN), where experiments are used as the main methods and execution time is used as the benchmark. Three areas of experiments are performed: automation, data recovery and data migration.

Cloud Storage Deployment Architecture is presented to demonstrate how this is designed and built based on group and research requirements. Selection of technology is explained and Cloud Storage is private-cloud SAN architecture made up of different NAS services. There are two premises: University of London Data Centre and GSTT/KCL. 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. There are arrows showing how automated and secure backups take place between Data Centre and GSTT/KCL.

Automation for Cloud storage has 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 automation in Cloud is more established than data recovery and data migration of single large files, and these two are perhaps challenges that data-intensive research need to overcome. Relative performance is between Cloud Storage and traditional storage have been presented and comparisons will be discussed.

Percentage of failure rate is calculated for backup automation, data recovery and data migration where backup automation stays below 2% of failure rate. The failure rate increases like an exponential manner up to 14.6% when data recovery volume increases up to 10,000 files. Similarly, an exponential increase is experienced when data migration increases up to 20.4% when data migration disk increases up to 1 TB per file. Our results demonstrate data recovery and data migration for thousands of files (that each has up to 1TB) have to be resolved and improved prior dealing with challenges in petabytes of storage. In summary, our main contributions include reduction of costs, time-saving to perform backups and user friendly interfaces.

References

Abu-Libdeh, H., Princehouse, L. and Weatherspoon, H., "RACS: A Case for Cloud Storage Diversity", SoCC '10 Proceedings of the 1st ACM symposium on Cloud computing, Indianapolis, Indiana, June 10-11, 2010.

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.

Beaty, K., Kochut, A. and Shaikh, H., "Desktop to Cloud Transformation Planning", "2009 IEEE International Symposium on Parallel and Distributed Processing", May 23-May 29 2009, Rome, Italy.

Bryant, R. E., "Data-Intensive Supercomputing: The Case for DISC", Technical paper, Carnegie Mellon University, October 2007.

Buyya, R., Yeo, C. S., Venugopal, S., Broberg, J. and 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, Volume 25, Issue 6, June 2009, Pages 559-616.

Buyya, R., Ranjan, R. and Calheiros, R. N., "InterCloud: Utility-Oriented Federation of Cloud Computing Environments for Scaling of Application Services", Algorithm and Architectures for Parallel Processing, Lecture Notes in Computer Science, 2010, Volume 6081/2010, 13-31 (Buyya et al., 2010 a).

Buyya, R., Beloglazov1, A., and Abawajy, J., "Energy-Efficient Management of Data Center Resources for Cloud Computing: A Vision, Architectural Elements, and Open Challenges", PDPTA'10 - The International Conference on Parallel and Distributed Processing Techniques and Applications, 12-15 July 2010, Las Vegas, USA (Buyya et al., 2010 b).

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.

Chang, V., Li, C. S., De Roure, D., Wills, G., Walters, R. and Chee, C. (2011) The Financial Clouds Review. International Journal of Cloud Applications and Computing, 1 (2). pp. 41-63. ISSN 2156-1834, eISSN 2156-1826 (Chang et al., 2011 a)

Chang, V., De Roure, D., Wills, G., Walters, R. and Barry, T. (2011) Organisational Sustainability Modelling for Return on Investment: Case Studies presented by a National Health Service (NHS) Trust UK. Journal of Computing and Information Technology, 19 (3). ISSN Print ISSN 1330-1136 | Online ISSN 1846-3908 (In Press) (Chang et al., 2011 b)

Chang, V., De Roure, D., Wills, G. and Walters, R. (2011) Case Studies and Organisational Sustainability Modelling presented by Cloud Computing Business Framework, International Journal of Web Services Research. ISSN 1545-7362 (In Press) (Chang et al., 2011 c)

Chang, V., Wills, G. and Walters, R. (2011) Towards Business Integration as a Service 2.0 (BIaaS 2.0), In: IEEE International Conference on e-Business Engineering, The 3rd International Workshop on Cloud Services - Platform Accelerating e-Business, 19-21 October, 2011, Beijing, China. (Chang et al., 2011 d).

Chang, V., Wills, G. and Walters, R. (2011) The positive impacts offered by Healthcare Cloud and 3D Bioinformatics. In: 10th e-Science All Hands Meeting 2011, 26-29 September 2011, York (Chang et al., 2011 e).

Chang, V., Wills, G., Walters, R. and Currie, W. (2011) Towards a structured Cloud ROI: The University of Southampton cost-saving and user satisfaction case studies. Sustainable Green Computing: Practices, Methodologies and Technologies (Chang et al., 2011 f).

Hey, A. J. G., "The fourth paradigm: data-intensive scientific discovery", Microsoft Publication, 2009, ISBN-10: 0982544200.

Latch, E. K., Dharmarajan, G., Glaubitz, J. C. and 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, DOI 10.1007/s10592-005-9098-1, Springer 2006.

Moore, R. W., Baru, C., Marciano, R., Rajasekar, A. and Wan, M., "Data-Intensive Computing", Chapter 5, Book chapter of "The Grid: Blueprint for a New Computing Infrastructure", ISBN 1558609334, 1999.

Moretti, C., Bulosan, J., Thain, D., and Flynn, P.J., "All-Pairs: An Abstraction for Data-Intensive Cloud Computing", IEEE International Symposium on Parallel and Distributed Processing, 2008, IPDPS 2008, 14-18 April 2008, Miami, USA.

Vo, H. T., Chen, C. and Ooi, B. C., "Towards Elastic Transactional Cloud Storage with Range Query Support", Proceedings of the VLDB Endowment, Volume 3 Issue 1-2, September 2010.