A case study on data protection and security decisions in cloud HPC

Morgan Eldred, Dr. Alice Good and Dr. Carl Adams
School of Computing, University of Portsmouth, Portsmouth, U.K.
morgan.eldred@myport.ac.uk, alice.good@port.ac.uk carl.adams@port.ac.uk,

Keywords: CLOUD COMPUTING, HIGH PERFORMANCE COMPUTING, SECURITY FRAMEWORK, DATA PROTECTION

Abstract: This paper reports on a case study that was conducted in a cloud HPC, one that used very sensitive and confidential data. The study aimed to explore the security challenges and practicalities that occur within a cloud HPC project and to develop a method for making critical security decisions. Action research was used to examine the nuances throughout the project as the service was moved from on-premise into a public cloud HPC, lasting over one year from start to finish. The study was able to identify some emergent issues affecting initiation, technical security challenges and the evaluation of a significant change in a HPC provisioning model.

1 INTRODUCTION

During the last 20 years there has been a continuing trend towards IT industrialisation and commoditization. This has resulted in IT services becoming repeatable and reusable by a broad range of both end-users and service provider, through advancements in technology and the accelerated growth in use of the Internet. These factors constitute the basis of a discontinuity that offers opportunities to shape the relationship between those who consume and those who provide IT services.

Cloud services provide a new way of delivering computing resources. Several types of cloud computing platforms exist, of which the main types are public, private and hybrid. Public clouds are normally offered by commercial organisations that provide access for a fee. Private clouds exist within are contained within a specific organisation and typically are not available for outside use. Hybrid clouds are a mixture of private and public clouds with the typical setup being that of a private cloud that has the ability to call upon additional resources from a public cloud (Chang, 2014).

Cloud HPC are computer clusters located in the cloud which address complex computational requirements, support applications with significant processing time requirements, or require processing of significant amounts of data. (C. Vecchiola, et al, 2009). These solutions in many cases do not use virtualization, but rather an architecture based around docker containers and the unique method that the technology has in sharing resources.

The main advantage of cloud computing is the ability of equilibrating the access to computing resources for all types of businesses, regardless their dimensions and investment capabilities. These advantages include cost efficiency, scalability, concentration, security and accessibility with a further list below. The advantages of cloud HPC is that they offer the potential to enhance innovation, as users can benefit from the emerging HPC ecosystem that is industry-agnostic and one that continually upgrades. This allows for organisations to quickly try new scientific approaches, with the elasticity of scaling up and down the infrastructure as required (Gartner 2015).

This paper outlines the overview, key issues and themes that emerged in a study of a large scale project within a mid-sized multinational company that ran a pilot to provision a scientific simulation software package via a public cloud.

2 RESEARCH METHODOLOGY

The research was conducted via an action research approach, through a case study that used an iterative

approach for collecting and analysing data. Benefits of this approach are that the research focused on practical problems, formulating solutions through the empowerment of the research to engage within the research and the subsequently through the execution activities (Mayer, 2000). Action research methodology typically takes a five step approach, as follows:

- Step 1: Identify the Problem
- Step 2: Devise a Plan
- Step 3: Act to Implement a Plan
- Step 4: Observe
- Step 5: Reflect and Share

Using this methodology, the approach started with identifying the problem, what are the security decisions that occur during a cloud HPC project. The second step was then to devise a plan around how to capture the security architecture and challenges that occur during a cloud HPC project. The next step was to execute the plan and execute the project. This part of the approach is where the action research takes place via the iterative approach. After the plan was implemented, the researcher observed if the outcomes had been achieved. Once the observation was completed, the researcher then reflected upon the entirety of the research and times the whole research approach may start over again (McCallister, 2011).

The researcher was a participant observer during the case study and was present for: top management meetings; this included, starting from project inception, all the way throughout the whole project. The access provided to the research provided rare insight into what the security decisions are during a large scale cloud HPC project. The value of the research was that it used an academic approach to a real-world case study.

The research design itself used a mix of deductive, qualitative and an inductive approach. Quantitative methods were used to determine the technical success of the project and what security decision points occurred. Interviews were conducted among a sample frame of 24 members of the organisation to understand and to identify key themes and issues that arose and to identify specific areas for the security decisions. The list of interviewee's consisted of those within the project team and stakeholders of the project, along with those whom would also be interested in the results of the project. A large portion of the data collected was related to the security decisions, and the key themes and issues that occurred, focusing deeper into the security decisions, challenges that occurred within a cloud HPC.

3 CASE STUDY

The case study is based on a midsized international company, with average revenues of between $8-10 billion dollars, located over seven countries and four continents.

The company invested in a formal project to explore the possibility of migrating on premise HPC that was used to run scientific simulations to a cloud HPC. This was due to the benefits of moving from on-premise to a cloud provisioning model were that it would enable the scientific community within the company to flexibly increase compute via a cost effective, on-demand, pay-per use model (Jackson et al, 2012). Cloud HPC are also increasing emerging as scalable and pay-on-demand solutions (Gartner, 2015).

The project itself was a multimillion dollar project that lasted over a year and consisted of a five person project team, along with twelve other stakeholders whom were involved in the project. The data used within the scientific simulations was extremely confidential and a major aspect was to determine the decisions points required for moving to a cloud HPC.

If the project was successful, the new capability from cloud HPC would enable to the company to compete with competitors that invested in large scale on-premise super-computing environments.

3.1 Problem Statement

The organisation was looking to become innovative as the cost of in-house supercomputing was extremely expensive, and it was not able to compete with the largest companies within its sector. This led to the organisation creating a formal project to explore if cloud HPC would enhance the capability of work that scientists did when running simulations. To become a leading player, the firm was challenged with a need for superior simulation modelling, as both the supply of information and the sophistication of quantitative techniques increases, it required a vastly higher resolution of raw data that would generating unprecedented volumes of data. All this additional data enables finer-scale simulation. This is the reason cloud HPC was identified as a potential solution. It has been identified that several key challenges will moderate the adoption of cloud HPC (Gartner 2015).

- Network costs may significantly increase
- Rigid software licensing from independent software vendors that provide niche, complex and expensive solutions
- Customization ability
- Data security and transport issues

- Architecture concerns around the need to support multiple flavours of on premise and cloud HPC

Security has been identified as a concern for some organisations in the adoption of cloud, with privacy and data ownership amongst the key factors for organisations deciding not to move to the cloud (Chang, 2011), the common reason for security being the sceptical question “who would trust their essential data out there somewhere?” (Ambrust, 2010).

However due to the sensitive nature of the sensitive scientific data, the project would need to determine real practical questions in relation to the security decisions required in moving the service to a cloud HPC. Certain industries involved in scientific work, such as the petrochemical industry have very strict rules about the movement of data outside of their jurisdiction, due to the fact that the data is considered a national asset.

It has been identified that there is a direct link between data protection and the growth of cloud computing as numerous hypotheses have indicated that the scale of cloud computing implementations may be affected in the short to medium term due in part to data protection risks (Eldred et al, 2015).

Listed below are the risk related to this project.

- Data security: is affected by shared technology vulnerabilities, data loss and data leakage, hijacking of traffic, malicious insiders or insecure APIs, and the reverse threat model where by malicious customers are able to undermine cloud services (Lambo, 2012).
- Confidentiality issues; due to the possibility of data exposure and leakages, this may lead to confidentiality breaches of the sensitive scientific data.
- Data integrity issues; given the loss of control of data processes, the organisation might be faced with the compromising of the sensitive scientific data.
- Business continuity; the organisation may suffer from business discontinuities and concerns related to the reliability of the cloud HPC. The organisation would need to make sure that proper storage, backup, and disaster recovery systems are in place at the cloud provider premises (Lambo, 2012).
- Federated clouds; the above mentioned risks are heightened when using federated clouds that call for multiple data transfers and thus generate a higher loss of control over the sensitive scientific data.

3.3 Architecture

Taking the security requirements as highlighted in the problem statement along with considerations from the five related risks categories, the design architecture guiding principle was that the architecture needed to be secure, lean and agile, as the sensitivity of the data was critical, and that the resource intensive simulation would need to be conducted in an efficient manner. The architecture was designed to be reliable, elastic with the ability to dynamically scale up or down compute clusters as and when needed. The architecture simulated a real-life enterprise network within a cloud HPC scenario. This was to design the cloud HPC service in an almost identical fashion to that of the on-premise service. This architecture would then easily be able to efficiently move data in and out. A Virtual Private Cloud-VPC in the Amazon European datacentre was configured and setup to act as the enterprise network, with another VPC in an Amazon US datacentre was configured and setup to act as the cloud HPC. Connections between the two VPC’s was facilitated through the use of an OpenVPN.

A major design requirement for the cloud HPC was the ability to transfer large datasets between the enterprise network and the cloud HPC. This was achieved through a cloud network attached storage-NAS server that was provisioned in the cloud, with a virtual device in the Amazon cloud configured to acts as a NAS front end to Amazon’s object based data cloud, simple storage service-S3. The large storage requirements, required the need for a common internet file system that is a standard way for users to share files across enterprise intranets and the internet, with a network file system interface. This design is commonly referred to as a cloud storage security gateway system, and is relatively known for being a secure way for encrypting and decrypting data via Amazon’s S3, as it examines the consistency of the contents during transmission and prevents data tampering (Wang et al 2013).

The next design step was to create the scientific simulation software head node in the cloud. This node was configured to be static and would be the node that scientific simulations would be submitted to and then it would dynamically-create a compute cluster.

A major security challenge in the architecture was the need to connect the scientific simulation software's physical Universal Serial Bus-USB license dongle to a virtual server. The architecture resolved this challenge by using a USB network device server that was placed within a de-militarised zone-DMZ within the organisations network and not the simulated enterprise network. This enabled the mapping of a USB port to a virtual server over the network. The USB port on the device server was then mapped to the scientific simulation license server in the DMZ and was configured with a public internet protocol address. The DMZ was as designed to enable traffic flow between the Amazon and license server.

A key component in the design was the NAS secure storage gateway. which was within the amazon cloud. The data was de/encrypted as it passed through the cloud NAS and resided within the amazon storage component, as it passed through the NAS and into the enterprise office cloud, where it was then de/encrypted and passed into the main data source.

4 ANALYSIS & DISCUSSION

After the architecture was designed and implemented based upon the security considerations, twenty four individuals involved in the project completed a questionnaire on the success of the project and some key trends and themes that emerged, with a further 10 being interviewed via open ended questions on a topic around the major security decisions that would need to be taken to move to cloud HPC, that came up during the project.

The 24 interviewees were asked a series of questions. As it relates to this paper, the critical questions focused on validating if the project was a success, as listed in Table 1. For data analysis purposes yes equating to a score of 1, while a no equated to a score of 0. The response and standard deviation were calculated as indicated in Table 1. Surprisingly only 25% indicated that the organisation was ready for cloud HPC.

TABLE 1: SUCCESS CRITERIA

Question	**Response**		**MEAN**	**STANDARD DEVIATION**
	Yes & (Total %)	**No**		
Is cloud HPC viable	17 (71%)	3	**0.850**	**0.366**
Is cloud HPC scalable	19 (79%)	4	**0.826**	**0.388**
Is the organisation ready for cloud HPC	6 (25%)	17	**0.261**	**0.449**

This line of questioning was included in this paper to indicate that the project was a success. To dive deeper into the security aspects, 10 selected members form the sample of 24, were then interviewed via open-ended questions to determine what critical decisions related to data protection and security occurred during the project. The information processed from the completion of the open ended interviews, resulted in the five security risk categories (Data Security, Confidentiality, Data integrity, Business continuity, federated clouds) which were then further developed into a new set of categories that covered these aspects along with governance and was used to develop the decision framework indicated in Table 2. categorized was developed. This decision framework was split into six three categories:

- Data confidentiality- the need to identify the sensitivity of the data
- Data storage- the need to understand how the data is stored and transported
- Disaster recovery- the need to look at how the organization would deal with impacts to any service outages
- Governance- the need to identify how cloud HPC will be governed and controlled

TABLE 2: CLOUD HPC SECURITY DECISION FRAMEWORK

Risk	Questions
Data Confidentiality	• What is the level of data confidentiality • How is unauthorized access controlled
Data storage and availability	• Where is the data stored? • Is the cloud solution federated? • Is the data encrypted?
Disaster Recovery	• What is the business impact in the event of an outage? • What are the back-up and failover plans?
Governance	• What is the process for the control and monitoring of moving solutions to Cloud HPC conducted? • Do auditing processes, exist?

The research shows that the evaluation and adoption of cloud HPC is a valid option for companies wanting to move scientific simulation software to the cloud. However many security considerations need to be addressed and not all scientific applications will be able to be moved to a cloud HPC. Other considerations will be that if a move to a cloud HPC is done, an organization will need to put in place the proper governance model, in ensuring that the control and monitoring of the solution can be done, along with the auditing processes. The outcome of the project, resulted in the organisation indicating it was not ready for cloud HPC. This was obtained via a questionnaire and was further built upon during the interviews where governance was highlighted as significant risk area, something which was not foreseen in the initial research or hypothesis.

5 CONCLUSION

The research involved an iterative methodology based upon the action research and covered all the stages of the cloud HPC project from initiation to evaluation. Cloud HPC is emerging, while cloud itself is maturing, but there is still a large amount of uncertainty that remains within the adoption for enterprises. Specifics include the organizational changes that are needed, along with the security, legal and privacy issues that cloud computing raises (A. Khajeh-Hosseini, 2010).

Cloud HPC is increasing emerging as an attractive offering for business that are required to run scientific computing. Due to the sensitive nature of scientific data, specific decisions around managing the security aspects are increasingly becoming a priority for organizations in understanding if they can transition sensitive data from in-house supercomputing clusters to cloud HPC. This paper has reported on research exploring the practicalities of conducting a significant cloud HPC project, providing a reference architecture for securing the a cloud HPC running scientific simulation software, along with providing a decision framework, that can assist information technology managers, whom may want to undertake a cloud HPC project. The research shows that the evaluation and adoption of cloud HPC projects may have considerable change to business practices, such as the need to introduce governance policies and checks.

I. FUTURE WORK

This research provides a security decision framework to address critical questions that need to be addressed when moving sensitive data and scientific applications to a cloud HPC. Further insights could be developed around building metrics around the decisions in the framework. Other aspects would be to further develop what an organisation requires to be ready for cloud, and how with this be managed and controlled. The research provided in this paper can provide a good foundation towards better understanding the wider impact that cloud HPC will have, however it can be enhanced by building out a cloud HPC framework for understanding the intracicies of moving to a HPC cloud and data in and out of the HPC cloud.

REFERENCES

M. Armbrust, A. Fox, R. Griffith, A. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, M. Zaharia, *A View of Cloud Computing*, Vol. 53 No. 4 ed. , ACM, 2010.

V. Chang, *The Business Intelligence As a Service in the Cloud*, 37, 512-534 ed. , Future Generation Computer Systems, 2014.

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

M. Eldred, R. Mcavey, *Prepare for a Qunatum Shift in Upstream Modeling*, Gartner, 2015

M. Eldred, C. Adams, A. Good, *Impact of EU data protection laws on cloud-computing: Capturing cloud-computing challenges and fault lines*, IGI, 2015

A. Khajeh-Hosseini, I. Sommerville, I. Sriram, *Research Challenges for Enterprise Cloud Computing*, 1st ACM Symposium on Cloud Computing, 2010

T. Lambo, Why You Need a Cloud Rating Score, CloudSecurityAlliance, January 2012, retrieved from: https://cloudsecurityalliance.org/wp-content/uploads/2012/02/Taiye_Lambo_CloudScore.pdf

J. McCallister, *Contemporary Social Work Issues*, MSW, 2011

J. Meyer, *Using qualitative methods in health related action research*, 320: 178-181 ed. , British Medical Journal, 2000

C. Vecchiola, S. Pandey, R. Buyya, *High-Performance Cloud Computing: A View of Scientific Applications*, Pervasive Systems, Algorithms, and Networks, 10th International Symposium on Pervasive Systems, and Networks, 2009

H.F. Wang, L.J. Wang, Pingjian Institute of Information Engineering, Beijing, China, Chinese Academy of Sciences, 2013