Abstract

Cloud computing provides access to on-demand computing resources and storage space, whereby applications and data are hosted with data centers managed by third parties, on a pay-per-use price model. This allows organizations to focus on core business goals instead of managing in-house IT infrastructure.

However, as more business critical applications and data are moved to the cloud, service availability is becoming a growing concern. A number of recent cloud service disruptions have questioned the reliability of cloud environments to host business critical applications and data. The impact of these disruptions varies, but, in most cases, there are financial losses and damaged reputation among consumers.

This thesis aims to investigate the threats to service availability in cloud computing and to provide some best practices to mitigate some of these threats. As a result, we identified eight categories of threats. They include, in no particular order: power outage, hardware failure, cyber-attack, configuration error, software bug, human error, administrative or legal dispute, and network dependency. A number of systematic mitigation techniques to ensure constant availability of service by cloud providers were identified. In addition, practices that can be applied by cloud customers and users of cloud services, to improve service availability, were presented.

Keywords: service availability, high availability, service failure, service disruption, cloud computing, cloud outage, denial of service, security, disaster recovery.
Acknowledgements

Life can only be understood backwards but it must be lived forward. There is no beginning without an end but the end is programmed right into the beginning. The researchers of this project work has always believed that the learning process should involve a mastery of certain fundamentals, we stand even at top of the most high hill, just to see a far behind us, we mean a far beyond our imaginative thinking behold all what we see marvels us.

All we could see was a great pile of pages in their mountain. Indeed, the list of those who have in one way or the other contributed to our three years of academic learning. We cannot but say a very big thank you to all. First and foremost, our thanks go to Almighty God who spares our lives up till today. For the knowledge and the ability to follow through with our undergraduate studies, we give thanks and adoration.

We would like to express our profound gratitude to our able and dynamic lecturers and thesis supervisor and our programme director in person of Ola Flygt for guiding us through successful completion of the thesis with suggestions and feedback. We are very much thankful to the examiner of the course- Mathias Hedenborg for his support throughout the research.

We would also like to thank the respondents who have contributed towards the survey part of this thesis. Finally, we are greatly thankful to our beloved parents for their moral and financial support they had given us to reach our goals.

Yours truly,

Adekunle Adegoke
Emmanuel Osimosu
Contents

1 Introduction ................................................. 1
  1.1 Background ............................................. 1
  1.2 Problem Definition, Aims and Objectives ................. 2
    1.2.1 Problem Definition ............................... 2
    1.2.2 Aims ............................................. 3
    1.2.3 Objectives ....................................... 3
  1.3 Research Methodology .................................. 3
  1.4 Thesis Outline ........................................ 4

2 Computer Security Overview ............................... 6
  2.1 Meaning of Computer Security ........................... 6
  2.2 Main Goals of Computer Security ....................... 6
    2.2.1 Confidentiality .................................. 7
    2.2.2 Integrity ........................................ 7
    2.2.3 Availability ..................................... 7
  2.3 Types of Information Security ......................... 7
    2.3.1 Data Security .................................... 8
    2.3.2 Application Security ............................ 9
    2.3.3 Network Security ................................ 10
    2.3.4 Operating System Security ....................... 10
  2.4 Information Security Standards ......................... 12

3 Cloud Computing Overview ............................... 14
  3.1 Definition of Cloud Computing .......................... 14
  3.2 Essential Characteristics of Cloud Computing .......... 15
  3.3 Service Models of Cloud Computing ...................... 15
  3.4 Deployment Models of Cloud Computing .................. 16
  3.5 Roles .................................................. 16
  3.6 Benefits of Cloud Computing ........................... 17

4 Service Availability Issues in Cloud Computing .......... 19
  4.1 What is Service Availability ........................... 19
  4.2 What Service Availability Means in Era of Cloud Computing 20
  4.3 Measuring Availability ................................. 20
  4.4 Cost of Unavailability ................................ 22
    4.4.1 Costs of lost business .......................... 24
    4.4.2 Cost of lost reputation .......................... 24
    4.4.3 Cost of penalties ................................ 24
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.4.4</td>
<td>Cost of lost worker hours</td>
<td>24</td>
</tr>
<tr>
<td>4.4.5</td>
<td>Recovery Cost</td>
<td>25</td>
</tr>
<tr>
<td>4.5</td>
<td>Threats Classification</td>
<td>25</td>
</tr>
<tr>
<td>4.5.1</td>
<td>Power Outages</td>
<td>25</td>
</tr>
<tr>
<td>4.5.2</td>
<td>Hardware Failures</td>
<td>26</td>
</tr>
<tr>
<td>4.5.3</td>
<td>Cyber Attacks</td>
<td>27</td>
</tr>
<tr>
<td>4.5.4</td>
<td>Configuration Error</td>
<td>30</td>
</tr>
<tr>
<td>4.5.5</td>
<td>Software bug</td>
<td>30</td>
</tr>
<tr>
<td>4.5.6</td>
<td>Human Error</td>
<td>31</td>
</tr>
<tr>
<td>4.5.7</td>
<td>Administrative or Legal Disputes</td>
<td>32</td>
</tr>
<tr>
<td>4.5.8</td>
<td>Network Dependency</td>
<td>32</td>
</tr>
<tr>
<td>5</td>
<td>Testing for Service Availability in Cloud Computing</td>
<td>33</td>
</tr>
<tr>
<td>5.1</td>
<td>Limitations</td>
<td>33</td>
</tr>
<tr>
<td>5.2</td>
<td>Objectives</td>
<td>33</td>
</tr>
<tr>
<td>5.3</td>
<td>Related Work</td>
<td>34</td>
</tr>
<tr>
<td>5.4</td>
<td>Cloud Providers and Technologies</td>
<td>34</td>
</tr>
<tr>
<td>5.4.1</td>
<td>Cloud Service Providers and Platforms</td>
<td>35</td>
</tr>
<tr>
<td>5.4.2</td>
<td>Programming Languages and Libraries</td>
<td>35</td>
</tr>
<tr>
<td>5.4.3</td>
<td>Web Servers</td>
<td>36</td>
</tr>
<tr>
<td>5.5</td>
<td>Implementation</td>
<td>36</td>
</tr>
<tr>
<td>5.5.1</td>
<td>Building the Test Application</td>
<td>36</td>
</tr>
<tr>
<td>5.5.2</td>
<td>API Configuration</td>
<td>37</td>
</tr>
<tr>
<td>5.5.3</td>
<td>Integration</td>
<td>38</td>
</tr>
<tr>
<td>5.5.4</td>
<td>Server Latency</td>
<td>38</td>
</tr>
<tr>
<td>5.6</td>
<td>Result</td>
<td>38</td>
</tr>
<tr>
<td>6</td>
<td>Survey on Service Availability Issues in Cloud Computing</td>
<td>41</td>
</tr>
<tr>
<td>6.1</td>
<td>Rationale for Survey</td>
<td>41</td>
</tr>
<tr>
<td>6.2</td>
<td>Source of Data Collection</td>
<td>41</td>
</tr>
<tr>
<td>6.3</td>
<td>Survey Questions Formulation</td>
<td>41</td>
</tr>
<tr>
<td>6.4</td>
<td>Analysis and Result of Survey</td>
<td>42</td>
</tr>
<tr>
<td>7</td>
<td>Best Practices for Service Availability in Cloud Computing</td>
<td>44</td>
</tr>
<tr>
<td>7.1</td>
<td>A Shared Responsibility</td>
<td>44</td>
</tr>
<tr>
<td>7.2</td>
<td>What Cloud Providers Should do to Ensure Service Availability</td>
<td>44</td>
</tr>
<tr>
<td>7.2.1</td>
<td>Geographic Redundancy</td>
<td>45</td>
</tr>
<tr>
<td>7.2.2</td>
<td>Network Redundancy</td>
<td>46</td>
</tr>
<tr>
<td>7.2.3</td>
<td>Power Redundancy</td>
<td>46</td>
</tr>
<tr>
<td>7.2.4</td>
<td>Hardware Redundancy</td>
<td>47</td>
</tr>
<tr>
<td>7.2.5</td>
<td>Recovering From Service Failures</td>
<td>48</td>
</tr>
</tbody>
</table>
7.3 What Cloud Customers Should do to Improve Service Availability
7.3.1 Design For Failure ............................................. 49
7.3.2 Determine Which Applications are Mission Critical ...... 50
7.3.3 Multiple Cloud Providers ..................................... 50
7.3.4 Cloud Monitoring .............................................. 51

7.4 What End users of Cloud Services Should do to Improve Service Availability ................................................. 51
7.4.1 Uninterrupted Internet Access ................................. 51
7.4.2 Adaptability and Learning .................................... 52
7.4.3 Feedback Report ................................................ 52
7.4.4 Password Management ....................................... 52

8 Limitations, Conclusion and Future Work 53
8.1 Thesis Limitations ................................................ 53
8.2 Conclusion ......................................................... 53
8.3 Future Work ....................................................... 54

References 62

Abbreviations 64
List of Figures

1.1 Challenges/Issues of Cloud Computing (Source = [7]) . . . . . 2
4.1 Availability of Service in Cloud Computing . . . . . . . . . . . 21
4.2 Availability of Service in Traditional IT Infrastructure . . . . . 21
4.3 Measuring Availability . . . . . . . . . . . . . . . . . . . . . . 23
5.1 Structure of the Test Application . . . . . . . . . . . . . . . . 37
7.1 Cloud Computing Security Players and Roles . . . . . . . . . . 45
7.2 Relationship Among End users, Providers and ISPs . . . . . . 47

List of Tables

4.1 Availability Percentage and Downtime Allowed . . . . . . . . . . 23
4.2 Downtime Statistics of Top Cloud Service Providers from 2007
to 2012 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 24
4.3 Total and Average Downtime For Each Service Provider and
Their Economic Impact . . . . . . . . . . . . . . . . . . . . . . . . 26
5.1 Availability Percentage and Downtime of the Deployed
Applications . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 39
1 Introduction

This chapter aims to introduce the thesis to the reader. It begins with giving a background on the research problem and significance of the problem. The chapter continues with sections stating the problem definition, aims and objectives of the thesis, followed by limitations and specific methods used in the study. In the last section, the structure of the report is outlined.

1.1 Background

The adoption of cloud computing has grown exponentially over the last few years. One third of all Internet users allegedly visit sites hosted on Amazon AWS cloud infrastructure daily and around one percent of Internet traffic in North America is reportedly served by Amazon AWS. According to the International Data Corporation (IDC), Western Europe’s cloud market is expected to grow to 15 billion euros by 2015, from 3.3 billion euros in 2010, translating to a 35% growth [1]. Cloud computing provides on-demand computing resources and storage space on a pay-per-use price model. An obvious benefit of this approach is cost effectiveness: eliminates the need to purchase and maintain in-house servers, scalability: allocated computing resources and storage space can increase or reduce depending on the demand.

Despite the benefits, there are potential risks. Between LinkedIn, Twitter and Yahoo password breaches [2] [3] [4], news of cloud security incidents have grabbed headlines in recent months. However, it is not only data breaches security professionals should be worried about, but equally availability breaches. Take for instance, the Amazon AWS cloud outage that occurred in April 2012, which took down many major websites including Reddit, Foursquare, Quora and Hootsuite [5]. This single, widely reported incident brought attention to the availability risks involved in cloud computing. While there were no reports of data breaches, the incident underscores the essentiality of continuous availability of some cloud services.

According to a survey conducted by IDC among 263 security professionals in 2009, the number one concern of organizations moving to cloud computing is security, while the second top concern is availability (Figure 1.1). However, because of the increasing migration of mission critical applications and sensitive data to cloud environments, cloud availability is no longer solely associated with accessibility but security as well [6]. After all, you cannot guarantee your data or application on the cloud is safe if you do not have access to them. That is, availability must be
first attained, before confidentiality and integrity can be achieved. As more enterprises adopt cloud computing, it is the responsibility of cloud service providers to ensure constant availability of critical business cloud services and data. In addition, cloud consumers can complement this effort through a number of techniques subsequently discussed in this report.

1.2 Problem Definition, Aims and Objectives

This section aims to discuss the problem definition of the thesis, its aims and objectives. It will give a clear detail of the research goals.

1.2.1 Problem Definition

Despite the fact that cloud computing has become ubiquitous and is gradually becoming technology option for most enterprise IT organizations around the world, there are numerous challenges causing drawbacks which needs to be dealt with. Data privacy and integrity is very paramount to business continuity using cloud solutions but there is one aspect of cloud computing services that remains an issue, and this is availability.

The threat against service availability in cloud computing encompasses lack of innovative ideas to ensure constant service delivery, outages problems, lack of operational resilience and lack of infrastructure robustness. These are the major problems definition to be addressed in this research.
1.2.2 Aims

The aim of this thesis is to investigate the threats to service availability in cloud computing and to provide some best practices to mitigate some of these threats.

1.2.3 Objectives

The following objectives have been set out in order to achieve the aims of the thesis.

- To understand cloud computing, what it is, how it works, and the benefits provided by this technology.
- To understand how services are accessed in cloud computing compared to traditional IT infrastructure.
- To perform a first-hand investigation to test the service availability goals of three major cloud service providers.
- To use survey to investigate threats faced by cloud service providers in their journey to ensure high availability of service. In addition, the steps taken to achieve this goal.
- Investigate publicly reported cloud service outages and disruptions from 2008 to date, identifying the causes of the failures.
- To investigate how cloud service providers can ensure constant availability of service.
- To investigate how customers of cloud service providers, and users of cloud services can improve availability of service.

1.3 Research Methodology

The thesis work will start off with collecting information generally on cloud computing, the main idea and focus on availability issues related to services offered by the cloud service providers. Selected research papers already published by large IT consortium like ACM, IEEE, large database of Google Scholar will be reviewed to get the current trend of the focus of the project. This information will be used to formulate the initial findings in the theoretical part of the work.

Another method we will use is qualitative survey techniques whereby a survey is prepared asking concerns questions on service availability in the
cloud from different cloud service providers in the Nordic region. We will do the follow up by sending reminders and making telephone call to get many respondents as much as possible.

Subsequently we will implement a thorough testing of service availability of major cloud players such as Microsoft Windows Azure, Amazon Web Service and Google App Engine by deploying a web application to these cloud platforms for a fixed period of time, collecting the percentage of uptime and comparing it with an expected uptime value. The result of such test can be used to evaluate cloud service providers who provide Infrastructure as a Service (IaaS), Platform as a Service (PaaS) or applications on the cloud with Software as a Service (SaaS) model.

Finally the result of survey, testing implementation result and our research findings of some related work that shows the issue of cloud service availability will be put together to summarize the best practices to achieve cloud computing services availability. The work was mainly conducted in Linnaeus University with a supervisor from Department of Computer Science, the survey will be sent out to Cloud services providers in Sweden and Finland and we will make use extensively of research papers from ACM, IEEE and other organizations.

1.4 Thesis Outline

This thesis is structured into the following chapters:

**Chapter 2 - Computer Security Overview** introduces the reader to computer security concepts such as: the meaning of computer security and its main goals. In addition, types of information security and information security standards.

**Chapter 3 - Cloud Computing Overview** provides an overview of cloud computing. Essential characteristics, service models, and deployment models of cloud computing are reviewed. Additionally, benefits provided by cloud computing are briefly discussed.

**Chapter 4 - Service Availability Issues in Cloud Computing** examines the threats to service availability in cloud computing. Service availability concepts in general, and in cloud computing are also discussed. In addition, the chapter discusses how to measure availability, and the cost of unavailability.

**Chapter 5 - Survey on Service Availability Issues in Cloud Computing** conducts a survey among cloud service providers, regarding service availability issues they face in cloud computing.

**Chapter 6 - Best Practices for Service Availability in Cloud Computing** investigates how cloud service providers can ensure availability
of cloud services. In addition, how cloud consumers or customers and end users of cloud services can improve availability of cloud services.

Chapter 7 - Conclusion and Future Work discusses the thesis findings and suggests possible supplement or additional work relating to the thesis work.
2 Computer Security Overview

The security concerns that organizations faces in cloud computing are no different from the ones they face in traditional IT infrastructure. The difference is that the security controls are now delegated between cloud customer and cloud service provider. To understand the security and availability challenges faced in cloud computing, it is essential to possess a fundamental understanding of computer security.

This chapter will discuss the fundamental concept of security, the focus is on computer security with its main goals whereby three main goals will be discussed in-depth and how they affect each other to achieve security in computing. It continues with a section that list types of computer security, its meaning and where they are applied. The last section in the chapter focuses on information security standards and policies that an organization conforms to which is demonstrated in a methodical and certifiable manner to achieve practices and procedures.

2.1 Meaning of Computer Security

When can we say that computer is secure? How do we verify the claims that this computer is immune against any forms of possible threats? These are all the questions that comes to mind when discussing about computer security. So in the wider sense, we might say a computer is secure if it is free from interference through any attacks and safe from threats, and computer security is the discipline that helps in protecting these assets and make us less worry about attacks on our computers. In summary when we are talking about computer security, we are focusing on three important aspects of any computer-related system: confidentiality, integrity and availability. According to National Institute of Standards and Technology (NIST), Computer Security is defined as:

"The protection afforded to an automated information system in order to attain the applicable objectives of preserving the integrity, availability and confidentiality of information system resources (included, hardware, software, firmware, information/data, and telecommunications)”, [8].

2.2 Main Goals of Computer Security

The term computer security has different meanings which are based on what different generations of computers describe. In earlier generations of computers, security is specialized in protecting the facility where computer
core engines was considered to be safe from any form of destruction, as technology in computing continue to move higher, users beginning to focus its security on protecting data and its validity. Considering all these principles computer was created, today computer security are built on three main goals which they are:

- Confidentiality
- Integrity
- Availability

2.2.1 Confidentiality
Confidentiality is when data is concealed from all but only accessible to those that have permission to use it. Confidentiality guarantees that the computer resources can only be accessed by the authorized people. The word access that is used in this context is to describe any act of whether viewing, printing or having the knowledge of the information by reading it [9].

2.2.2 Integrity
When an information is trustworthy, which means it has not been changed anyhow whether it is deliberate or by accident. It serves the purpose for which the data was created for then we say the data has integrity. Integrity of data is when its state does not change from when the last authorized users use it while data is available when the authorized user can access it in a preferable format as expected at that particular time needed [9].

2.2.3 Availability
Availability is the main focus of this research work within the context of cloud computing. When information is readily available and accessible by the users that have authorization to see it and it comes at the appropriate time then we would say the data has availability guarantee [9].

2.3 Types of Information Security
Most computer security expert categorized types of computer security into two major types; they are Software and Hardware security with a number of categories within them. We will discuss major types of computer security in this section but before we go into details, generally software security can be described as system security from viruses, worms, malicious programs
and spyware while hardware security can be regarded as security of physical
devices like mainframes, storage devices and external portable memory to
mention a few. Also Software security could be referred to ways an attack
could be launched on data streams and software without any direct contact
of physical devices or hardware.

Most common types of computer security that will be discussed in this
section are data security- keeping data safe from corruption and its access
is controlled, application security- measures taken during the life-cycle of
an application such as analysis, design, development, testing, deployment,
and maintenance, network security- protection of data during transmission
across the networks and lastly operating system security is the process of
ensuring operating system integrity, confidentiality and availability. These
are measures used to protect operating system from the threat of attacks
by viruses, worms, malware or hacker’s intrusions thereby allows different
applications and programs to perform required tasks and halt unauthorized
interference [10].

2.3.1 Data Security

Data in computer systems are always vulnerable to threats and that is why
it is one of the most popular types of computer security. Data security is
maintained when the three main objectives of computer security is achieved,
they are integrity, confidentiality and availability. Making data secure is the
objectives of so many computer systems and a complete solution to data
security must meet those objectives. According to Dorothy and Robling,
data security is:

"the science and study of methods of protecting data in
computer and communication systems from unauthorized

The current world we live in with growing rapid technology and the
heavy dependence on Internet for our daily transactions and existence is
growing like never before [12]. Today, we have the ability to utilize scalable,
distributed computing environments within the confines of the Internet, a
term known as cloud computing. This primary distributed computing
platform has created another security issues in IT industries which rise for
concern on the issue of data security in the cloud. Major players in cloud
computing business such as Google, Yahoo, Microsoft, and Amazon are
developing day to day security mechanisms to ensure data integrity,
confidentiality and availability of their client data situated on their
infrastructures. Data stored in the cloud could be ordinary that can be
found in any public source or private such as financial records, credit card numbers, medical records, shipping manifest for hazardous materials. These different categories pose a security concern for cloud service providers and how it is to be managed, who should be responsible or is it going to be a shared responsibility between the provider and clients. These create security loopholes and special interest from security professionals on what they should address in data security when it comes to cloud.

While cloud computing has made it possible to allow startups who cannot afford large computer infrastructures to use services over the cloud for their businesses, it has also become crucial target for cyber criminals that once an attack could get through cloud services providers endpoint then numerous of websites could be compromised. Legal interpretation of all these security concerns, who should take responsibility in case of an attack, is it the owner of the data or the cloud service provider? How do we verify actual endpoint where the attacks took place? Is it at the client side or provider side? All these questions are subjected to computer security main goals which if sustainable cloud data security is to be implemented then confidentiality, integrity and availability must be ensured [13].

2.3.2 Application Security

Application security is not just adding password protected login screen when deploying an applications. Security frameworks need to be put in place when building applications, there are patterns that developers need to follow that will address the security issues in the application development. These patterns constitute to the guidelines that developers can follow when building applications which is going to help them. Single Access Point limits application entry point to one single point in which users are not allowed to get through a back door to access sensitive data. Check Point pattern handles different security breaches and provides punishment for the security violation. Roles define what the users can do and cannot do while Session distributes information about the user throughout the application. Limited View presented users with legal options and Secure Access Layer pattern is used by applications to communicate with external systems in a secure manner. All these patterns work together to provide necessary security within an application, the main challenge in the application security is what security policies to be followed and correctly coded so that there would be high assurance that the hacker cannot bypass it [14].

Under cloud computing perspective, applications security are mainly connected to one of the service models in cloud computing (SaaS), Software as a Service. The security issues are tightly connected to the interface
where those applications are accessed which is web browser. Web applications to be hosted on cloud infrastructure are validated and scanned for any vulnerability using web application scanners. Firewalls that are discovering any vulnerability such as examining HTTP requests/responses for applications specific vulnerabilities are steps to ensure applications security in the cloud. According to OWASP [15], who listed ten most critical web applications vulnerabilities amongst are injection, cross site scripting (Input validation) weaknesses. Applications security misconfiguration with multi-tenancy where each tenant has their own security configurations that may conflict with each other will create security holes, it is highly recommended to depend on cloud provider security controls to enforce and manage security in a dynamic and robust manner[16].

2.3.3 Network Security

Network security is an important area of computer security; nowadays everything is interconnected via networks which give reasons for network security to grow to protect these assets from perceived threats or potential attacks. Network security encompasses all aspect of computer security which depends on all the cryptographic tools available, agile program development processes, operating system controls, trust and evaluation and assurance methods, and inference and aggregation controls [17]. According to SANS Institute, a leading organization in the world in computer security training, Network Security is defined:

"the process of taking physical and software preventative measures to protect the underlying networking infrastructure from unauthorized access, misuse, malfunction, modification, destruction, or improper disclosure, thereby creating a secure platform for computers, users and programs to perform their permitted critical functions within a secure environment", [17].

While this definition encompasses network security in computer security as a whole, the forefront of network security in cloud computing lies in virtual machine security, secure communication of hypervisors and protection against external attacks.

2.3.4 Operating System Security

An operating system is the controlling brain-box of computer system with two goals: it controls shared access and implementing an interface to allow
that access. The two major goals has many different support activities including identification and authentication, naming, filing objects, scheduling, communication among processes, and reclaiming and reusing objects. Each of these activities has security implications from the simple ones supporting single task at a time to complex multi-user and multitasking systems. There are many commercially operating systems but we draw examples largely from two families: Microsoft Windows NT, 2000, XP, 2003 server and Vista operating systems and Unix, Linux and their derivatives called Unix+. We have mobile operating systems such Apple IOS and Android, in which all of them required different security level and functions [9].

The perspective of this section will focus more solely on cloud operating system and its security issues as the research work is on availability issues around cloud computing services. The recent deployment of Google’s Chrome OS, an open-source cloud based operating system, has added more worries to concerns for the security of cloud computing especially when the most tasks are handled outside of the user’s hardware and control. Virtualization is the central technology which makes cloud computing possible whereby a single PC or server simultaneously run more than one session of operating system. Virtual Machines also known as Hypervisors handles communication between the different operating systems and the CPU within the server, storage of data and network connection.

This is the basis of Cloud computing technology explored by one of the Cloud giant-Amazon to create IaaS (Infrastructure as a Service) for companies around the world. With this new technology innovation, concerns grow more in security of operating system running simultaneously with the help of virtualization technologies. Several attacks are possible such as Wrapper attacks and DOS attacks. Wrapper attack could let the attacker inject duplicate a fragment of XML while adding additional code that would lead the computer to do additional unwanted tasks [18].

Google created a cloud operating system called Google Chrome OS which uses browser for its input/output operations as the main source of connectivity with the cloud services. There are many issues facing security for browsers within the cloud. The first common line of defense for browsers is for servers to use the Same Origin Policy (SOP) which is for the server to monitor the original location of the browser when the request was made and only accept requests if the request comes from the same location. Nowadays the major security protection for browsers is using Transport Layer Security (TLS), however, the major flaw to TLS is phishing which is where the users are tricked by a malicious website to access the confidential information of the users such as login information and if this happen at any
endpoint on the cloud then this is a major security incident to the entire architecture. Operating system security has changed its tune from just traditional computing now to cloud which make it more important and sensitive because if taken control of is possible then the entire cloud space could be under attack.

### 2.4 Information Security Standards

Information security standards can be regarded as what organizations laid down to measure the quality control of their best practices and procedures. Standards in general ensure accepted characteristics of products and services for example like safety, reliability, efficiency and interchangeability. These are accepted regulations to implement information security controls to meet an organization’s requirement. The International Organization for Standardization (ISO) is an international body composed of representatives from different national standards organizations. It is the largest body comprises of voluntary International Standards.

ISO has member bodies which are national bodies consider to be the most representative standards body in each country, correspondent members (countries that do not have their own standard organization) and subscriber members (countries with small economies). ISO/IEC 27000-series comprises of information security standards was published together the two bodies namely; International Organization for Standardization and International Electrotechnical Commission (IEC). This series is responsible for all organization of different sizes covering privacy, confidentiality and IT or technical security issues. There are currently four published standards in the series I mentioned above: they are 27001, 27002, 27005 and 27006.

The 27001 standards sets out the steps required for an organization’s Information Security Management Systems (ISMS) to achieve certification. ISO/IEC 27002 lists security control objectives and recommends a range of specific security controls. It contains best practices and security controls in different areas of information security management such as security policy, organization of information security and human resources security. Another published information security standard is ISO/IEC 27005 which provides guidelines for information security risk management. Organizations such as enterprises for profit making, government bodies and agencies, non-profit organizations that intend to manage risks that could compromise the organizations’ information security considered the usage of ISO/IEC 27005 series. The 27006 standard specify the steps of certification and registration processes that must be followed by the certifying bodies. Some standards
are still underdevelopment which would be ready in the future with the current demands of guidelines in IT industries [19].
3 Cloud Computing Overview

The chapter aims to provide a brief overview of cloud computing. Section 3.1 to section 3.5 provides the standardized definition of cloud computing, its essential characteristics, service and deployment models respectively. Section 3.6 highlights some benefits provided by cloud computing.

3.1 Definition of Cloud Computing

There are a number of proposed definitions of cloud computing, however, definitions alone are unintuitive. We attempt to explain what cloud computing is using real world examples.

Take for instance a small or medium business enterprise where resources such as emails, applications or databases are hosted on local servers. To maintain the servers and keep things running smoothly, a team of skilled network engineers and administrators is needed. This is where cloud computing comes into play. Cloud computing allows you to outsource part or your entire IT infrastructure to data centers managed by third parties, thereby cutting off the need of purchasing and maintaining local servers. The main task of these third-parties data center providers is to maintain, secure and update service equipment allowing you to focus on developing your business rather than managing it. This model allows for a lot of flexibility and is cost-effective since you only have to pay for only as much capacity as is needed or consumed.

Similarly, from an end-user perspective, cloud computing allows storing, processing and managing data on networked servers hosted by third parties. Google Drive and Dropbox are two popular examples of cloud storage services. Dropbox, however, stores files given to them on Amazon’s Simple Storage Service (S3) [20], rather than storing them on its local servers which brings us back to the first example.

The National Institute of Standards and Technology (NIST) defines cloud computing as follows:

"Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g. networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction", [21].
3.2 Essential Characteristics of Cloud Computing

According to NIST, the following are the five essential characteristics of the cloud model:

- **On-demand self-service.** This refers to the ability of a customer to provision computing capabilities or services whenever needed without any interaction with the cloud service provider or a support desk [21].

- **Broad network access.** This refers to the ability to access cloud services or technologies from a wide range of client platforms, anytime, anywhere in the world, over the internet. That is, access to services is platform independent [21].

- **Resource pooling.** This refers to the idea of sharing, where the physical resources and logical resources of the service provider are pooled together and are shared by multiple customers. Resources pooling is accomplished by the concept of virtualization [21].

- **Rapid Elasticity.** This refers to the ability to very quickly provision resources and at the same time have the ability to scale down the amount of resources used. [21].

- **Measured Service.** This refers to the ability to monitor resources used and very accurately determine the amount of resource used which allows customers to only pay for only consumed resources. [21].

For any cloud technology to be considered a true cloud technology, it must meet each of these five characteristics.

3.3 Service Models of Cloud Computing

According to NIST, the cloud model is composed of the following three service models:

- **Software as a Service (SaaS).** This is a model whereby a cloud service provider allows a consumer to access their online applications via client interface, in most cases, a browser. Consumers have very little control over the backend infrastructure like the operating system the application is running on, server, network, and the like. Things of such are taken care of by the provider. The consumer however, retain control over their data. [21].
• **Platform as a Service (PaaS).** This model is common among developers as it provides a platform to deploy applications to. A platform refers to a deployment environment on the web. A PaaS provider provides everything needed to run an application deployed to its platform, such as, application server, operating system, database management system, and the like [21].

• **Infrastructure as a Service (IaaS).** This model allows a consumer to retain most control over the cloud environment. The cloud service provider provides infrastructure such as data storage capabilities, networking activities, server hardwares and the like. The consumer retains the control over what operating system to deploy on it and other configurations. [21].

The models are collectively referred to as SPI model and as observed, as a consumer gains more control, more responsibilities are inherited and vice versa.

### 3.4 Deployment Models of Cloud Computing

According to NIST, the cloud model is composed of the following four deployment models:

- **Private cloud.** A private cloud is dedicated to a single enterprise, organization or entity. This can be managed by the organization itself or a cloud service provider, on or off premises [21].

- **Community cloud.** A community cloud is a private cloud shared among several entities. Organizations that have shared interest or concerns such as policy, security requirements, and interoperability maybe share the same cloud infrastructure, managed on or off promises by the organizations or a third party. [21].

- **Public cloud.** A public cloud offers services to the general public. Cloud infrastructure and resources are shared among consumers [21].

- **Hybrid cloud.** In a hybrid cloud model, the organization may use any combination of the other cloud models. [21].

### 3.5 Roles

The following are the five major participating actors identified by NIST [22].
- **Cloud consumer** - An individual or organization that uses cloud services and products.

- **Cloud Provider** - An individual or organization that provides services to cloud cloud customers.

- **Cloud Broker** - A third person facilitator whose aim is to negotiate cloud offerings and services with a cloud provider, on behalf of a cloud customer.

- **Cloud Auditor** - A third person agent who evaluates the security and performance of cloud services.

- **Cloud Carrier** - An organization responsible for transporting of cloud services between cloud providers and cloud consumers through e.g network access.

### 3.6 Benefits of Cloud Computing

Figure 3.1 shows a graph that represents the results of a survey conducted by the International Data Corporation (IDC) in 2008 amongst 244 IT executives and chief information officers, regarding the benefits of cloud computing. The respondents clearly sees cloud computing as a solution to their need for rapid deployment in their various organizations.

The next top three benefits deals with cost-effectiveness through a pay-as-you-use model, reducing the need to hire staff to manage in-house IT infrastructures and a monthly payment that lowers cost barriers for new start-ups.
**Figure 3.1:** Benefits of Cloud Computing (Source = [7])
4 Service Availability Issues in Cloud Computing

As more enterprises move their business applications and data to the cloud, service availability is becoming a growing concern. Assuring uninterrupted availability of cloud services is still a challenging issue for cloud service providers. A number of recent cloud outages have questioned the reliability of cloud environments to run mission critical applications. Amazon Web Services (AWS) was down for 49 minutes on January 13, 2013, and reportedly lost close to 5 million US dollars according to [23]. Visits to Amazon.com, which runs on AWS cloud computing platform, returned 503 errors which suggest possible denial of service attack [24]. Other major cloud service providers such as Google and Windows Azure have had their share of outages too.

This chapter examines threats to availability of cloud services and proposes a classification of the threats, but first, we will briefly discuss what service availability means in general, in cloud computing, how it is measured and the costs of unavailability.

4.1 What is Service Availability

The term service availability has been used multiple times in the report with assumption that the meaning is obvious to the reader. This section aims to make sure that we have the same understanding of what service availability means to the reader. Following a definition by the Service Availability Forum, service availability is:

"an extension of high availability, referring to services that are available regardless of hardware, software or user fault and importance", [25].

A key requirement of mission critical systems is that they must maintain uninterrupted service even in the event of hardware or software failures, that is, they must be fault tolerant. SAF identified four key principles to achieve this requirement. They are as follows: [25]:

- Redundancy.
- Stateful and seamless recovery from failures
- Minimizing mean time to repair (MTTR).
- Monitoring.
These principles not only encompasses reactions to failure, but also proactive monitoring whereby actions are taken before failure occurs. These principles applies to cloud computing services and will be subsequently discussed in the report.

4.2 What Service Availability Means in Era of Cloud Computing

Availability is a crucial factor in any IT infrastructure, since loss of availability means loss of production. News of cloud outages has grabbed headlines in recent times, and in most cases, there are huge financial losses. Nonetheless, how does service availability in a cloud environment compares to service availability in traditional IT infrastructures? Failures are inevitable in a system, why is it such a big deal in cloud computing? Has loss of service availability in cloud computing been greatly exaggerated despite cloud’s proven record of success? The answer lies in the mode of service delivery of each infrastructure. Figure 4.1 and figure 4.2 illustrates an example of how services are accessed in the cloud and in traditional IT infrastructure.

Figure 4.2 depicts a scenario whereby services are hosted centrally on a cloud provider’s infrastructure and accessed via the Internet by end users. An obvious advantage of this approach is for instance, a centrally-managed security updates and hardware upgrade. Figure 4.1 depicts a scenario whereby services are hosted locally on each organization’s IT infrastructure and also accessed over the Internet by end users. Services hosted on cloud platforms include database, block storage, object storage, runtime, compute, network and more. Considering that a cloud service provider like AWS hosts several millions of applications and over two trillion data objects [26], a complete outage of service means denial of service to the hosted applications and data which could devastate the global IT industry. In this regard, a cloud infrastructure can be considered as one big single point of failure. In chapter 7, we will look at how such issues can be addressed with techniques such as redundancy and multi-data center replication.

4.3 Measuring Availability

Before we begin to address the issue of service availability in cloud computing, we need a way to measure availability. Typically, availability refers to the percentage of time a system is up, that is, a system with 100% availability is always up [27]. Abbadi expressed in [28], the concept of availability as:

$$\text{Availability} = \frac{MTBF}{(MTBF + MTTR)} \times 100$$
Figure 4.1: Availability of Service in Cloud Computing

Figure 4.2: Availability of Service in Traditional IT Infrastructure

Where,
MTBF refers to the *mean time between failures*, and MTTR refers to the *mean time to repair*.

Figure 4.3 illustrates various metrics used to calculate availability. Take for instance, a situation whereby a cloud customer has experienced two hours of downtime during one month, here is how to calculate the availability for that month:

\[ MTBF = 30 \text{ days} \times 24 \text{ hours} \times 60 \text{ minutes} = 43200 \text{ minutes} \]

\[ MTTR = 2 \text{ hours} \times 60 \text{ minutes} = 120 \text{ minutes} \]

Recall,

\[ Availability = \frac{MTBF}{MTBF + MTTR} \times 100 \]

Therefore,

\[ Availability = \left( \frac{43200}{43200 + 120} \right) \times 100 \approx 99.72\% \]

Similarly, suppose a cloud service provider guarantees 99% uptime a year, this offer sounds great, except it could mean 3 days downtime a year which is terrible for mission critical applications. Here’s how to calculate the allowable downtime in a year:

\[ Total \ \text{uptime in a year} = 12 \text{ months} \times 30 \text{ days} \times 24 \text{ hours} \times 60 \text{ minutes} = 518400 \text{ minutes} \]

\[ The \ \text{Cloud Provider’s MTBF in a year} = 99\% \]

\[ Allowable \ MTTR \ in \ a \ year \ (percentage) = 1\% \]

Therefore,

\[ Allowable \ MTTR \ in \ a \ year = \left( 518400 \times 1\% \right)/100 = 3.65 \text{ days} \]

Table 4.1 lists the approximate amount of downtime allowed for a cloud service provider to achieve a certain percentage of availability.

The International Working Group on Cloud Computing Resiliency (IWQCR) provides a summary of downtime statistics of some top cloud service providers from 2007 - 2012 [29] as shown in table 4.2. According to [30], approximately 165,000 websites went down for a period of one week due to NaviSite outage in 2007. A failed data center migration was cited as the cause.

### 4.4 Cost of Unavailability

Unavailability of cloud services can result in loss of productivity, revenue and reputation. The total impact of the downtime is calculated by summing up all the losses incurred during downtime. The following factors should be considered depending on the services offered [31].
Figure 4.3: Measuring Availability

<table>
<thead>
<tr>
<th>Uptime(%)</th>
<th>Downtime(%)</th>
<th>Downtime Per Year</th>
<th>Downtime Per Week</th>
</tr>
</thead>
<tbody>
<tr>
<td>98</td>
<td>2</td>
<td>7.3 days</td>
<td>3 hrs, 22 mins</td>
</tr>
<tr>
<td>99</td>
<td>1</td>
<td>3.65 days</td>
<td>1 hr, 41 mins</td>
</tr>
<tr>
<td>99.8</td>
<td>0.2</td>
<td>17 hrs, 14 minutes</td>
<td>20 mins, 10 secs</td>
</tr>
<tr>
<td>99.9</td>
<td>0.1</td>
<td>8 hrs, 45 mins</td>
<td>10 mins, 5 secs</td>
</tr>
<tr>
<td>99.99</td>
<td>0.01</td>
<td>52.5 mins</td>
<td>1 mins</td>
</tr>
<tr>
<td>99.999</td>
<td>0.001</td>
<td>5.25 mins</td>
<td>6 secs</td>
</tr>
<tr>
<td>99.9999</td>
<td>0.0001</td>
<td>3.15 secs</td>
<td>0.6 secs</td>
</tr>
</tbody>
</table>

Table 4.1: Availability Percentage and Downtime Allowed
Table 4.2: Downtime Statistics of Top Cloud Service Providers from 2007 to 2012

<table>
<thead>
<tr>
<th></th>
<th>2007 (hr)</th>
<th>2008 (hr)</th>
<th>2009 (hr)</th>
<th>2010 (hr)</th>
<th>2011 (hr)</th>
<th>2012 (hr)</th>
<th>Total (hr)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amadeus</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Facebook</td>
<td></td>
<td>1</td>
<td>3</td>
<td>4</td>
<td></td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>ServerBeach</td>
<td>4</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>Paypal</td>
<td></td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>Google</td>
<td>4</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>Yahoo!</td>
<td></td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>6</td>
</tr>
<tr>
<td>Twitter</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>7</td>
</tr>
<tr>
<td>Amazon</td>
<td>24</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>24</td>
</tr>
<tr>
<td>Microsoft</td>
<td>24</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>31</td>
</tr>
<tr>
<td>Hostway</td>
<td>72</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>73</td>
</tr>
<tr>
<td>BlackBerry</td>
<td></td>
<td></td>
<td>72</td>
<td></td>
<td></td>
<td></td>
<td>72</td>
</tr>
<tr>
<td>NaviSite</td>
<td>168</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>169</td>
</tr>
<tr>
<td>OVH</td>
<td>168</td>
<td></td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td>170</td>
</tr>
</tbody>
</table>

4.4.1 Costs of lost business

This refers to the direct financial loss incurred as a result of a downtime. Take for instance, a trading or finance system where a few minutes of outage can translate to several millions of kronor in impact. Network World editor Brandon Butler wrote in [32], "Amazon.com’s latest earnings report showed that the company makes about $10.8 billion per quarter, or about $118 million per day and $4.9 million per hour.” This means for every hour of downtime, a loss of almost $5 million is incurred by Amazon.com.

4.4.2 Cost of lost reputation

Outages can cause long term damage to a cloud service provider’s reputation. Reputation matters to consumers since choosing a provider is normally based on reputation-based trust model and availability evidence.

4.4.3 Cost of penalties

Some cloud service providers and customers face penalty fees in the event of an outage. This is especially the case when there is loss of confidential customer data.

4.4.4 Cost of lost worker hours

Loss of productivity due to lost worker hours can be costly to cloud service providers and customers in the long run. In most cases, permanent employees are paid regardless of whether they meet their work targets.
4.4.5 Recovery Cost

This refers to cost incurred to recover from an outage, such as, paying employees to work overtime, and disaster recovery services.

IWQCR released a summary of the total and average downtime of cloud service providers in table 4.2 with their economic impact [29], as shown table 4.3.

4.5 Threats Classification

Various incidents result in unavailability of cloud services. The aim of this section is to examine threats to service availability in cloud computing. We have identified eight categories of threats based on security incidents reports and press releases published from 2008 to date. The categories are as follows:

4.5.1 Power Outages

A power outage refers to the loss of electrical power in an area. The single biggest failure point of any cloud infrastructure is its power supply. A recent research conducted by RightScale found that 33 percent out of 27 cloud outages it investigated in 2012 were due to power loss or failed backup. Six of the power outages were however caused by Hurricane Sandy [33]. A power outage could be short-term or long-term and each both pose different impact from loss of revenue to loss of reputation. Long term outages can be caused by earthquakes, flood, fires and other natural or man-made disasters. We give some examples:

- VMware’s cloud computing platform CloudFoundry.org suffered service disruptions on April 25 and April 26, 2011. VMware official Dekel Tankel explained in [34], that power outage and subsequently a misconfiguration error, was the cause of the service failure.

- Cloud Computing giant Salesforce.com experienced a major outage on July 10, 2012. The outage reportedly lasted for 13 hours, leaving thousands of customers unable to use its Customer Relationship Management (CRM) service. The cloud service provider’s status page reported that power disruption was the cause of the outage [35].

- Web hosting and cloud service provider DreamHost experienced a temporary power systems failure at its data center in Irvine, California, on March 2013. The outage reportedly caused hours of downtime for more 350,000 customers [36].
<table>
<thead>
<tr>
<th>Service Provider</th>
<th>Total (Hour)</th>
<th>Average (Hour)</th>
<th>Availability</th>
<th>Cost/Hour (USD)</th>
<th>Cost (USD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amadeus</td>
<td>1</td>
<td>0.167</td>
<td>99.998%</td>
<td>89.00</td>
<td>89.00</td>
</tr>
<tr>
<td>Facebook</td>
<td>3</td>
<td>0.500</td>
<td>99.994%</td>
<td>200,000</td>
<td>600,000</td>
</tr>
<tr>
<td>ServerBeach</td>
<td>4</td>
<td>0.667</td>
<td>99.992%</td>
<td>100,000</td>
<td>400,000</td>
</tr>
<tr>
<td>Paypal</td>
<td>5</td>
<td>0.833</td>
<td>99.990%</td>
<td>225,000</td>
<td>1,125,000</td>
</tr>
<tr>
<td>Google</td>
<td>5</td>
<td>0.833</td>
<td>99.990%</td>
<td>200,000</td>
<td>1,000,000</td>
</tr>
<tr>
<td>Yahoo!</td>
<td>6</td>
<td>1.000</td>
<td>99.989%</td>
<td>200,000</td>
<td>1,200,000</td>
</tr>
<tr>
<td>Twitter</td>
<td>7</td>
<td>1.167</td>
<td>99.987%</td>
<td>200,000</td>
<td>1,400,000</td>
</tr>
<tr>
<td>Amazon</td>
<td>24</td>
<td>4.000</td>
<td>99.954%</td>
<td>180,000</td>
<td>4,320,000</td>
</tr>
<tr>
<td>Microsoft</td>
<td>31</td>
<td>5.167</td>
<td>99.941%</td>
<td>200,000</td>
<td>6,200,000</td>
</tr>
<tr>
<td>Hostway</td>
<td>72</td>
<td>12.000</td>
<td>99.863%</td>
<td>100,000</td>
<td>7,200,000</td>
</tr>
<tr>
<td>BlackBerry</td>
<td>72</td>
<td>12.000</td>
<td>99.863%</td>
<td>200,000</td>
<td>14,400,000</td>
</tr>
<tr>
<td>NaviSite</td>
<td>168</td>
<td>28.000</td>
<td>99.667%</td>
<td>100,000</td>
<td>17,000,000</td>
</tr>
<tr>
<td>OVH</td>
<td>170</td>
<td>28.333</td>
<td>99.667%</td>
<td>100,000</td>
<td>17,000,000</td>
</tr>
<tr>
<td>Total</td>
<td>568</td>
<td>94.667</td>
<td>99.917%</td>
<td>71,734,000</td>
<td>71,734,000</td>
</tr>
</tbody>
</table>

Table 4.3: Total and Average Downtime For Each Service Provider and Their Economic Impact

4.5.2 Hardware Failures

PC Mag defined hardware failure as:

"A fault within the electromechanical components or electronic circuits of a computer system. Recovery from a hardware failure requires repair or replacement of the offending part", [37].

Any kind of computer systems will in fact suffer hardware failure at one point or another. The key is to eliminate single point of failure through redundancy and clustering. Clustering provides high availability through intelligent combination of servers with the purpose of achieving minimal downtime. That is, if a one server in a two-cluster architecture goes down, the second server takes over [38]. Hardware failures without adequate recovery or restoration plan can lead to data loss resulting in loss of revenue and damaged reputation. There are numerous cases wheres losses of availability or data were as a result of hardware failures:

- Amazon’s European websites were down for half an hour on December 12, 2010. The downtime was initially thought to be caused by an attack but an Amazon spokesperson later confirmed it was due to hardware failure [39].

- Software as a Service (SaaS) provider Jive suffered a major outage on January 14, 2011, leaving close to 500 customers unable to serve their
wiki pages. The downtime was blamed on a hardware failure within a storage system [40].

4.5.3 Cyber Attacks

Cloud service providers can address the issues of power outages or hardware failures through redundancy of servers or data centers so that if a data center is affected by a disaster or there is a hardware failure, they can immediately switch to the redundant resources, thus avoiding downtime. Cyber-attacks, however, exploit the software layer of cloud systems which persist across servers and data centers [1]. This section aims to examine some of the prominent attacks against availability of service in cloud computing.

**Denial of Service (DoS) Attacks.** Following a definition from the Computer Emergency Response Team (CERT), a denial of service attack is

"characterized by an explicit attempt by attackers to prevent legitimate users of a service from using that service", [41].

DOS attacks are often attempted by flooding the target with large amounts of traffic with the purpose of exhausting the target system’s resources. Some security experts have argued that cloud computing is more vulnerable to DOS due to its shared, on-demand nature. Typically, cloud infrastructures consist of virtual environments that allow multiple customers to share the same physical infrastructure or hardware. With the growing adoption of cloud computing and a centralized data center for applications, the magnitude of damage of a DOS attack is much more severe. When a virtual server running on a cloud platform or infrastructure is flooded with high traffic, the cloud computing operating system tries to absorb the traffic by providing more computational power. Eventually, the hardware would be unable to hold the workload and subsequent requests will be denied, not only by the virtual server but all virtual servers since they all share the same underlying hardware resources. Thus, the attacker does not need to flood all virtual servers since the same goal can be achieved with a single entry point [42].

**Authentication Attacks.** Authentication attacks attempts to exploit the authentication process a system uses to verify an application, service or user’s identity. In cloud computing, application and data are accessed over the Internet outside the traditional network perimeter of the cloud
customer. There is need to protect applications and data deployed to cloud systems from unauthorized access, deletion or modification which can result in unavailability of service. While basic authentication mechanism such as the use of username and password has proved little protection, strong authentication has drawn little attention mainly due to implementation cost [43]. For instance, to deploy and manage an application to Google’s cloud computing platform known as app engine, one only needs a Google email and password. The following are some of the attacks against authentication exchange or credentials.

**Man-in-the-middle attacks.** An attacker places him or herself in the connection path between the customer’s client and cloud system during an authentication exchange. The attacker then attempts to authenticate by posing as the customer’s client to the cloud system and the cloud system to the customer’s client.

**Phishing attacks.** An attacker attempts to lure a customer into revealing their authentication credentials. Typically, an email message or website asks the customer to verify their account information as part of normal security checks conducted by the email or cloud service provider.

**Insider attacks.** A customer’s authentication credentials can be compromised by an employee working for the cloud service provider who may also be in charge of managing confidential customer data. This could be as revenge to the employer or for other reasons.

**Keylogger attacks.** The attacker places a malicious program on a customer client device in attempts to monitor and capture keystrokes typed during an authentication exchange. Hardware keylogging involves an attacker physically plugging a device such as USB to the client device without the knowledge of the customer.

**Cloud Malware Injection Attacks.** This attack works slightly differently depending on the cloud service model targeted. In SaaS or PaaS, an attack relies on injecting the target cloud system with a malicious application instance in an attempt to lure the cloud system into treating it as a valid instance and serving requests with it. In IaaS, an attacker attempts to inject the cloud system with a malicious virtual machine implementation and to subsequently trick the cloud system into treating the new malicious instance as a valid instance. If successful, requests are
directed to the malicious virtual machine and service availability is disrupted. A proposed countermeasure is making sure the cloud system performs virtual machine integrity check before directing requests to it [42].

**Side Channel Attacks.** An attacker takes advantage of the shared-tenancy nature of a cloud computing environment by placing a malicious virtual machine implementation in close proximity to a target virtual machine running on the same physical machine. A research conducted by computer scientists from Massachusetts Institute of Technology and University of California, San Diego found that it is feasible for an attacker to map a cloud infrastructure’s internal architecture and subsequently locate the domain of the target virtual machine. The attacker can then monitor metrics such as cache, keystroke timings in an attempt to extract data hosted on the same machine [44]. The Cloud Security Alliance released a report in 2010 listing this attack as one of the top threats in cloud computing [45].

Lastly, we give a number of examples:

- Music sharing platform SoundCloud suffered two days outage from October 4 to October 6 in 2011 due to an attack. In statement posted on its blog page, the company confirmed that the service outage was caused by a distributed denial-of-service attack [46].

  The difference between a DoS and a (distributed denial of service) DDOS attack is that a DoS attack floods its target from one computer and one Internet connection, while a DDOS attack floods its target from multiple computers and Internet connections distributed globally into what is called a botnet.

- In 2012, credit card processor Global Payments announced a security breach of its system that allowed hackers to obtain 1.5 million credit numbers. Why the affected customers weren’t liable for any unauthorized transactions as a result of the attack, the impact on the card processor was estimated to be $84.4 million dollars [47].

- On April 29, 2011, Sony Playstation gaming network was breached and subsequently went offline. Hackers got access to 70 million user account information including user names, passwords, emails and so forth. It is possible that credit card information were obtained [48].

- In December 2011, online marking company Epsilon was hacked and customer data of CapitalOne, US Bank, Citi, BestBuy, JPMorgan Chase, Walgreens and TiVo were accessed by hackers [49].
4.5.4 Configuration Error

Even though cloud environments are largely automated, configuration errors are one of the major causes of service failures. With the ever increasing complexity of cloud infrastructures, small misconfiguration either in the hardware or software installation can easily take the whole cloud system down. 2012 saw some of biggest IT companies brought to their knees by configuration errors. We give some examples:

- Microsoft’s cloud computing platform Azure suffered an outage in August 2012 leaving customers in Western Europe unable to use the service for more than two hours. In a message posted on its blog, the company’s general manager, Mike Heil, explained that the service interruption was caused by a mis-configured network device that disrupted traffic in its West Europe sub-region [50].

- Users of Google’s email service, GMail experienced widespread service unavailability in December 2012. The issue was tracked to a faulty load-balancing configuration change [51].

4.5.5 Software bug

A software bug is a flaw in a computer program that makes the program do what it’s not designed to do or not do what it’s designed to do [52]. Examples of undesired behaviours includes crashing, freezing and memory leaks. Majority of software bugs are as a result of human error in software design and implementation stages of development. Software should therefore be tested thoroughly or handed over to quality assurance team to make sure they are bug free before deployed. Cloud systems are not immune to software bugs and cloud service providers should exercise automatic software and driver updates as a preventive measure. We give some examples of service failures caused by bug in the cloud computing system.

- Microsoft’s Windows Azure suffered a 12 hour outage in its Chicago, Dublin and San Antonio data centers in March 2012. The outage was triggered by a leap year software bug that prevented systems from calculating valid dates for security certificates [53].

- Google storage service, Drive, were hit by a series of outages between March 11 and March 17 2013, leaving users unable to access their files and document on the file storage system. Google issued a report about
the incident on March 18 blaming it on a bug in their network control software [54].

4.5.6 Human Error

The subject of human error is complex and cannot be explained in few sentences. Dekker in [55] defined two fundamental ways of looking at human errors: The Old View or The Bad Apple Theory and the New View. The Old View maintains that:

- Complex systems fail as a result of unpredictable nature of people.
- Human errors cause accidents: and
- Failure is an unexpected phenomenon.

The New View however maintains that:

- Failure is not caused by human error. Human error is the effect, or symptom of deeper trouble.
- Human error is not random: and
- Human error is the starting point of an investigation, not the conclusion.

While the Bad Apple Theory is a cheap and straightforward approach of dealing with failures, it ignores the fact that human errors sometimes reveal the shortcomings of a system. Service outages caused by human error are common in cloud computing and are on the rise. We give some examples:

- Healthcare IT provider, Cerner, suffered a US-wide outage on July 23, 2013, that affected Cerner’s hospital and physician customers. In response to an inquiry from InformationWeek Healthcare, the company’s spokesperson confirmed the outage was caused by human error [56].

- Amazon Web Service experienced an outage on Christmas Eve of 2012 that affected many of its customers notably Netflix. The outage was reported to be caused by an accidental data deletion by an employee at Amazon [57].
4.5.7 Administrative or Legal Disputes

Administrative or legal dispute involving a cloud service provider or its customer can result in unavailability of service. There have been examples of cloud service providers shutting down unexpectedly leaving customers unable to access or move their data and applications. It is therefore essential that a customer considers the flexibility of moving their data and applications to another service provider in case of a shutdown, bankruptcy, acquisition or other reasons. We give a couple of examples:

- In 2012, file sharing site MegaUpload was shut down by the US government over copyright violation. The site was at the time the 13th most visited website in the world, with around 180 million registered users hosting over 25600 terabytes of data [58].

- In August 2008, online storage service provider The Linkup shut down after losing access to "unspecified amounts of customer data". It was not clear whether it was the data loss that prompted the shutdown of the service but The Linkup CEO Steve Iverson confirmed that at least 55% of customer data were safe at the time of shutdown [59].

4.5.8 Network Dependency

A conventional threat to the existence of cloud computing is Internet connectivity. Similar to power outages, the Internet could as well be considered a single failure point of any cloud infrastructure. There are scenarios where Internet access is limited or non-existent making it impossible for users to access data or applications. An example of such scenario is when flying with airlines without Internet connection or when out of range of a Wi-Fi network. Cloud computing’s dependency on the Internet means downtime is certain since even the best Internet service providers experience downtime at one point or the other. Cloud computing thus makes cloud providers and their customer dependent on the efficiency of their Internet carriers. Caching techniques aid offline access to static files but are largely unreliable.
5 Testing for Service Availability in Cloud Computing

The concept of availability testing is broad and relative to individual circumstances. Besides not having a well-documented procedure to perform this task properly, availability means different things to different people. Should timely access be considered as a key component when designing the test? What if a web application is availability but users from one network segment cannot access it? Is a web application available if it performs 90% of its functions but not the other 10%? For this test, testing for service availability means running a web application for a fixed period of time, collecting the percentage of uptime, and comparing it with an expected uptime value. The collected uptime values will be compared with the uptime values guaranteed by each cloud platforms the applications were deployed to, as state in their Service Level Agreements (SLA). SLA refers to a contract between a cloud consumer and cloud service provider that specifies the level and quality of service expected. SLAs facilitates the negotiation process between the two parties [60].

5.1 Limitations

This test entails building a monitoring webpage that computes the percentage of uptime of a web application deployed to three different cloud platforms, displaying them with a graph. It may also happen that a web application is available but the monitoring service provider is down. We attempt to solve this problem by setting up a redundant secondary monitoring service. Also, due to limited computing resources, the deployed web applications will be monitored in 5 minute intervals.

5.2 Objectives

If you depend on the cloud, you need a way to monitor it. Our approach is performing availability check on the cloud application. While some cloud service providers provides service health dashboards that report uptime stats of their services, a third-party or independent monitoring is a more favourable approach. The objectives of this test are follows:

- Deploy a dynamic web application to the platform of three cloud service providers.
- Compute the uptime percentage of the three applications for a certain period of time.
• Compare the uptime percentage of each deployment to the availability percentage offered by each provider, as stated in their SLAs.

• Investigate any disruption that occurs.

The result of such test can be used to evaluate cloud service providers who provide IaaS, PaaS or applications on the cloud with SaaS model.

5.3 Related Work

Cedexis, a company that specializes in measuring Internet health through browser instrumentation in order to optimize its customers’ use of clouds and CDNs, conducted a test that analyzed the availability and performance of nine cloud platforms from roughly 300 million end users perspectives across the world. The test was conducted over a period of 7 days in 2011 [61]. Key findings include:

• *Regional effects matter tremendously, and the difference between countries can be significant* [61].

• *The average time to complete an HTTP request and receive a response, worldwide, was 426.4 milliseconds. The average availability worldwide was 97.69%* [61].

While this test involves a larger number of participants in terms of cloud platforms and end users than the one conducted in this thesis, it lacks a key element which renders the result almost unreliable. This key element is time. Ideally, since most cloud services provider measures the availability of their offered services on a yearly basis, service availability testing of the services should be ran for a year, or at a minimum, a month, in order to gain accurate results. The test conducted in this thesis involves testing and analyzing the availability and performance of three cloud platforms for a period of one month.

5.4 Cloud Providers and Technologies

The following cloud computing platforms and technologies were used in the test:
5.4.1 Cloud Service Providers and Platforms

- **Google App Engine (GAE)** - Google App Engine allows developers to deploy web applications to Google’s infrastructure. It offers a pay-as-you-use model allowing business to pay for only resourced used, and a scale-on demand to meet increasing traffic demands. They offers a free quote up to a certain level of consumed resources [62]. The free tier provides enough computing resources to perform the test.

- **Amazon Web Services Elastic Beanstalk** - AWS Elastic beanstalk allows for quick and easy way to deploy web applications to the Amazon’s infrastructure and manage them. Elastic Beanstalk combines AWS services such as Amazon Elastic Cloud Compute, Amazon Simple Storage Service, Amazon Simple Notification Service and Elastic Load Balancing, to provide cost-effective, high available and scalable infrastructure to customers [63]. They also offer free usage tier for a year with no service restrictions, although a valid credit card is required to sign up [64].

- **Windows Azure Cloud Services** - A Microsoft’s cloud application platform for building, deploying and managing applications on Microsoft-managed data centers. It provides support for many different programming languages and support PaaS and IaaS service models. They offer a 3 month free trial with no service restrictions but a valid credit card is required to sign up [65].

5.4.2 Programming Languages and Libraries

- **Java Enterprise Edition (Servlets)** - A dynamic web application was written for the test using servlets. Servlets are Java platform framework for building web based applications. To generate a dynamic content for each request, we used the java.util.Date class to get the current time on the server, and also pragmatically extract the server name and version to the browser. More information about building Java web applications with servlets can be found at [66].

- **HyperText Markup Language (HTML)** - The monitoring page is written using HTML, a markup language for creating web pages displayed with a web browser

- **JavaScript** - A programming language used by web browsers to show interactive content.
• **Google Charts** - Provide tools to easily add interactive chart to a web page.

### 5.4.3 Web Servers

- **Apache Tomcat** - An open source Java servlet container for serving Java-based web applications. The test applications deployed to AWS Beanstalk and Windows Azure run on tomcat.

- **Jetty** - An open source Java-based HTTP server and Java servlet container. GAE provides PaaS model and supports Jetty by default which means we only have to focus on the developing and deploying the application and GAE takes care of the server installation and configuration.

### 5.5 Implementation

Rather than building a monitoring service from scratch, we decided to use the one by [www.uptimerobot.com](http://www.uptimerobot.com), an uptime monitoring service provider, by integrating their features into our monitoring webpage. They are well-known, reliable and provide features such as email and sms notifications. The service is also free and web applications will be monitored in 5 minute intervals.

The service works by requesting a website header, and returned status codes like "200-ok" indicating the web application was successfully loaded or 400+ and 500+ indicating the web application is not loading. Several more checks is performed in the next 3 minutes in order to make sure it is veritably down after which it then sends an alert [67].

A second monitoring service [www.statuscake.com](http://www.statuscake.com), is used for redundancy purpose.

#### 5.5.1 Building the Test Application

The test application is written in Java. Figure 5.1 shows the general structure of the application.

The image `/image/NoImg.gif` is used to calculate the latency of the server which is the time it takes to receive and process a request for the image.

The Java class `com.testapp3.servlets.Testapp3Servlet.java` is invoked when the default page is loaded and dynamic contents are generated and served back to the web browser.
The application is then deployed to GAE, AWS and Windows Azure and served at gaeappspot.appspot.com, awsdeployment.elasticbeanstalk.com and azuredeployment.cloudapp.net/Testapp3 respectively.

5.5.2 API Configuration

After signing up for the monitoring service, we need a way to fetch and compute the uptime percentage from the monitoring service provider. Luckily, they provide an API layer with which we can fetch these data but first we need an API key to do this. Two kinds of API keys can be generated:

- **Main API key** - This key is used to add/delete/edit the monitors from a programming interface. A monitor refers to a web application or web site that is being monitored. Needless to say, this key should be kept secret.

- **Monitor-specific keys** - This key is specific to each monitors and can be used to fetch uptime data for a particular monitor. The API documentation can be found at [68].

Main API key is u82120-b345f8a0dccccf80d5172f850
Google App Engine monitor key is m775757931

Figure 5.1: Structure of the Test Application
AWS Elastic Beanstalk monitor key is m775757932
Windows Azure monitor key is m775757934

5.5.3 Integration

To fetch the uptime percentage for 1, 7, 30 and 365 days, we passed the main API key and monitor-specific API key into a URL as follows:

http://api.uptimerobot.com/getMonitors?apiKey=u82120-b345f8a0dccccf80d5172f85&monitors=775757931-775757932-775757934&customUptimeRatio=1-7-30-365&format=json

We also specified the data-interchange format as JSON. XML is also supported. The data is then saved into a JavaScript array object and passed as an argument to a method that generates and displays the graph for each monitor.

5.5.4 Server Latency

Another feature we have included in the monitoring webpage is a server latency test. Latency refers how much times it takes a server to receive and process a request. This is done by requesting a small image placed in a public folder of the web application such as gaedeployment.appspot.com/images/NoImg.gif. A starttime variable is defined to store the currenttime just before the image is requested and once the image is loaded to the browser, the starttime is subtracted from a new currenttime to get the response time in milliseconds.

As a matter of fact, two factors affects latency: the distance of the server from the client and how much data is being transported [69], although the latter is of no concern in this case since the same image is placed on the three applications.

5.6 Result

Before we begin to analyse the result of the test, we will briefly review the uptime percentage/availability guaranteed by each cloud service provider used in the test.

- Microsoft’s Windows Azure cloud service offers 99.95% monthly uptime percentage for computing services which translates to 21.56 minutes per month [70]. Customers are entitled to service credits if this percentage is not met.
Table 5.1: Availability Percentage and Downtime of the Deployed Applications

<table>
<thead>
<tr>
<th>Cloud Service Providers</th>
<th>Uptime (%)</th>
<th>Downtime</th>
</tr>
</thead>
<tbody>
<tr>
<td>Windows Azure</td>
<td>95.88</td>
<td>1779.84 minutes or 29.66 hours</td>
</tr>
<tr>
<td>Amazon AWS Elastic Beanstalk</td>
<td>93.41</td>
<td>2846.88 minutes or 47.45 hours</td>
</tr>
<tr>
<td>Google App Engine</td>
<td>100</td>
<td>0</td>
</tr>
</tbody>
</table>

- Amazon AWS offers a yearly uptime percentage of 99.95% [71]. This means the service could be down for 4.38 hours in a single month and up for the rest of the running year, unlike Windows Azure that could only be down for 21.56 minutes in one month. Customers are also entitled to compensation in the event the service level is not met.

- Surprisingly, Google App Engine also promises 99.95% monthly uptime percentage with a clause for compensation if the level is not met [72].

Note: different uptime percentages are guaranteed for different services. For instance, while Windows Azure offers 99.95% for computing services, they offer only 99.9% for storage services. We are not interested in storage service levels since our application required no storage or database.

Table 5.1 shows the metrics attained after a month (30 days) of the test. When we saw these numbers, we knew something went wrong somewhere. After thorough investigation, it turned out that the monitoring service we used reported false downtime alerts and the applications were in fact up the entire time. We knew the applications were up because the redundant secondary monitoring service did not issue any downtime alert and reported that the applications were up 100%

"Regarding the false/positives experienced 2 weeks ago, it was our fault. We upgraded our node.js version, faced incompatibility issues with some code. And, had a network problem at a close time. It took us some time to identify the issues. Everything is fixed afterwards and we added multiple false/positive controls, network controls and more to prevent that from happening again.\textsuperscript{,} [73]."

The monitoring service is in fact an example of Software as a service (SaaS) model of cloud computing and the problem experienced during the test is an example of loss of availability. However, we are not entitled to compensation or bothered to ask for any since the service is free.

The URL to the monitoring web page is http://homepage.lnu.se/student/eo222bp/index.html.
Server Latency. The average response time of the Google App Engine server is 146.9 milliseconds, while the average response times for AWS Elastic Beanstalk and Windows Azure are 246.2 milliseconds and 308 milliseconds respectively. These numbers simply refers to how much time it took the servers to receive our requests plus how much time it took our client to receive the responses from the servers. The apparent explanation for the difference in latencies is: the more geographically distant the server is, the higher the latency and vice versa. Similarly, the higher the number of routers between a server and a client, the higher the latency and vice versa. Cloud service providers have begun tackling this problem by building data centers around the world. So for instance, if you are deploying an application to a cloud platform and you expect majority of your users to come from a specific country, you can demand that your application be deployed to the provider’s data center located in your country if available, if not, to a country close-by.
6 Survey on Service Availability Issues in Cloud Computing

The purpose of this chapter is to get the cloud services provider perspective on availability issues in cloud computing and the challenges they face to ameliorate the problems in order to satisfy their customers in the long run. In this study, we used qualitative survey techniques to gain an understanding of cloud computing availability in the provider’s viewpoint.

6.1 Rationale for Survey

This survey is part of the research methodologies adopted to gather necessary statistics on behaviors existing between cloud service providers and their clients. The focus is on availability and the best practices to reduce the downtime and ensure continuity of cloud services. We conducted this survey as part of the effort to substantiate our findings on the extent of availability of cloud computing services. We sent out survey to some Cloud providers in Sweden and Finland in total of 30 companies but we got response from 4 providers during this research work. To complement this research study, we also conducted a testing implementation that monitors the uptime percentage of application deployed to the platform of three different major players in cloud computing world. As part of the research methodologies, we used questionnaire to get valuable answers to how Cloud services providers in Nordic region are ensuring availability of cloud services they provide for their clients.

6.2 Source of Data Collection

The survey was created with Google Spreadsheet and sent electronically by email to the cloud services providers afterwards we issued reminders and eventually called them up to get more responses.

6.3 Survey Questions Formulation

The questionnaire was developed from the knowledge gain from systematic review of research papers, cloud availability incidents report, journal reviews and background knowledge in information security. The survey consists of ten research questions that address availability concerns. They are as follows;

- How many employees work in your organization?
- What cloud services do you currently offer?
• How would you rate your service availability readiness for cloud deployments?

• What is the top business driver for your service availability goals?

• What strategies does your company use to ensure service availability use?

• What is your service availability goal?

• What were the causes for outages (all outages) you experienced in 2012?

• How is liability issues addressed in the case of damages occurring from interruptions in service?

• Which disaster recovery solutions do you currently use?

• How often do you test your disaster recovery plan?

6.4 Analysis and Result of Survey

The findings are:

• The respondents offers IaaS, SaaS and PaaS cloud service models out of which one offers the three services models including cloud brokerage services, two offers IaaS and SaaS, one offers only PaaS.

• Two of the respondents rate availability exactly as all other security capabilities of their cloud systems while two rate availability higher above them.

• All the respondents are very keen to ensure the customers service availability goals and satisfaction while two of them in addition lay emphasis about avoiding productivity loss.

• All the respondents employs high availability clusters to ensure service availability, two of the respondents used disaster recovery as a service while two also depend on restoration from back up. One of the respondents employs a technique called parallel systems.

• All the respondents claimed that they offer less than 8hrs downtime or 99. 91- 99.99

• Two of the respondent said they suffered outages caused by cyber-attack and human errors in the year 2012.
• The respondents said liability in terms of data loss depends on the Service Level Agreement and provisionally the root cause of the event.

• One respondent said they test their disaster recovery plan 2-3 times a year.
7 Best Practices for Service Availability in Cloud Computing

This chapter is divided into four sections. The first section discusses key players that must work together to ensure service availability and their roles. The second section discusses a number of ways cloud service providers can ensure high availability and continuity of service. Similarly, the third and fourth sections outline some best practices cloud customers and end-users of cloud services can follow to improve availability of cloud services.

7.1 A Shared Responsibility

When organizations host applications or data in the cloud, some security controls are transferred to the cloud service provider. Such controls comprise network, physical, data and virtualization security. Physical security includes redundancy, high availability infrastructure, secure premises, fail-over capability, and maintenance operations. Network security entails access control and firewall systems to prevent attacks and unauthorized access to the network. Data security entails using encryption techniques to protect data from unauthorized access, modification or deletion. Virtualization security entails securing virtual machines or server instances running on cloud infrastructures.

For applications deployed to the cloud, a cloud customer is responsible for application security which encompasses measures taken throughout the design and maintenance stages of an application. Additionally, application layer security controls such as testing, software maintenance, change control are employed by the customer. Figure 7.1 illustrates cloud computing security players and their roles.

Ensuring availability of service is a shared responsibility of a cloud provider, customer and end user of cloud services. End users have limited or no role in securing cloud applications but they can employ a number of techniques to ensure continuous access to their data in the cloud.

7.2 What Cloud Providers Should do to Ensure Service Availability

There are a number of ways cloud service providers can ensure service availability. This section aims to provide some best practices.
7.2.1 Geographic Redundancy

Service failures can be caused by natural disaster, power outages, hardware failures, cyber-attacks, configuration error, software bug and human error. The best practice for mitigating failures that cannot be rapidly recovered on premises is by deploying redundant systems to a separate physical location. The practice is sometimes referred to as geo-redundancy [74]. While this practice would benefit the cloud customers by aiding business continuity in the event of an outage, configuring and maintaining a redundant geographically separated data center that have the same capacity and capabilities of its primary counterpart is expensive [75]. Business and regulatory requirements sometimes dictates the need for backup systems to ensure data protection and business continuity in the
event of a failure.

A common approach is maintaining a primary data center that serves traffic and a secondary data center for backup or emergency fail-over in the event of an outage. Data can be replicated from the primary data center to the secondary data center using a full or real-time synchronization. A full synchronization is based on intervals, ideally, within a 24-hour period. A real-time synchronization however reduces the load and time required for a full synchronization [76].

In some cases, it might be desirable to serve traffic from both primary and secondary data centers. This approach may improve service latency since traffic is served to users by which data center is in close proximity to them.

7.2.2 Network Redundancy

Typically, cloud service end users connect to cloud resources via a public Internet. However, Internet Service Providers (ISPs) supply the end users connection to the Internet. Cloud service providers on the other hand connect to ISPs to provide cloud services to end users. ISPs can be thought of as cloud carriers in this sense. Figure 7.2 illustrates the relationship among cloud end users, providers and ISPs.

A cloud service provider’s reliance on a single ISP poses the problem of single point of failure. A simple solution to prevent loss of service availability due to a network outage is through the use of multiple ISP or Internet carriers. Such problem occurred in February 2012 when Australia’s biggest ISP Telstra suffered a network outage that last for 30 minutes affecting several businesses across the country [77].

7.2.3 Power Redundancy

In a redundant power supply configuration, if the primary power supply fails, the redundant power supply takes over and keeps the servers and data center running, preventing downtime. Typically, each redundant power supply unit is capable of powering the entire data center and the switch between the power supplies should be automatic so as to prevent any service interruption. A standby power generator seamlessly switches on within seconds of a power outage. \textit{Uninterrupted power supplies} (UPS) also referred to as battery-backup, provides emergency power supply for a short period of time in the event that a redundant power supply is unavailable.
7.2.4 Hardware Redundancy

The traditional way to mitigate service failures caused by hardware failures is by switching service from the failed hardware component to a redundant component or by repairing or replacing the failed component and restarting the software [78]. Hardware components include disk drives, servers, switches and routers. A technique known as *redundant array of independent disks* (RAID) allows redundant storage of data on multiple hard disk so that if one drive fails, availability is not interrupted and data loss is prevented [79]. There are a number of different RAID types but they are a combination of the most commonly used types which are:

**RAID 0.** Data is split and stored in two or more hard drives so that when the data is accessed, it is pulled from all the disks it was saved on. This offers high performance but no redundancy since a failure in one of the disk would result in data loss. This is also known as *stripping.*
RAID 1. Data is mirror or copied to the redundant disks. That is, everything on the primary disk is copied on the redundant disks. An obvious advantage of this technique is redundancy since the exact copy of the data can be recovered if the primary disk fails. The disadvantage is that only 50% of the total storage space of the disks is used and low performance compared to RAID 0. This technique is also known as mirroring.

A similar technique known as redundant array of inexpensive servers (RAIS) involves setting up redundant servers that provide the same service so that in the event of a failure of one server, service availability is not affected [80]. Similar technique can be applied to switches, routers and other hardware components.

7.2.5 Recovering From Service Failures

Some business critical applications are obviously mission-critical - that is, they cannot afford to go offline for even a short period of time. Such applications include e-commerce shop and e-banking. On the other hand, the criticality of some applications may not be apparent. Some applications can most likely accept a certain level of downtime with little consequences to the cloud customers. As a first preventive measure, the criticality of each application hosted should be determined. The following metrics should be checked for each application:

Recovery Point Objective (RPO). Following a definition from Symantec, RPO is defined as:

"the maximum tolerable data loss that is acceptable in a disaster situation", [81].

How much data loss can be tolerated in the event of a failure? A day’s worth of business data? a month?

Recovery Time Objective (RTO). Xand defined RTP as:

"the maximum tolerable length of time that a computer, system, network, or application can be down after a failure or disaster occurs", [82].

What is the maximum tolerable period of disruption before the business activities starts to suffer? Needless to say, a lengthy RTO means catastrophic impact on the business.
**Recovery Resources.** This defines the resources needed to recover from failure. Such resources include employees, equipment, hardware and redundant power supply.

After checking the three metrics for each application, determine whether all business critical applications can be recovered in parallel with the resources available. If the recovery needs surpass the available resources, a cloud-based disaster recovery should be adopted. A cloud-based disaster recovery solution allows a cloud service provider to replicate back up data to an off-site recovery infrastructure managed by a third party that guarantees restoration in the event of a failure or disaster.

### 7.3 What Cloud Customers Should do to Improve Service Availability

Cloud service customers can employ the following techniques to improve availability of their cloud services.

#### 7.3.1 Design For Failure

Just as it is important to design cloud infrastructures and hardware components for failure, applications deployed to cloud systems should also be designed for failure. Following an Amazon AWS cloud outage that took down Reddit, Quora.com, Scvngr.com, EveryBlock.com and GroupMe.com in early 2011, EveryBlock’s Paul Smith wrote on the company’s blog:

> "While the acute problem originated with AWS, EveryBlock is not without blame for this downtime. Frankly, we screwed up. AWS explicitly advises that developers should design a site’s architecture so that it is resilient to occasional failures and outages such as what occurred yesterday, and we did not follow that advice. As the team member who is de facto in charge of this area of our project, I apologize to our users who rely on EveryBlock to connect with their neighbors and get their local news", [83].

There are many different systematic approaches to designing cloud applications for failure, but the rule of thumb should always be: assume failures will occur, assume there will be cloud outages, assume your application will be stroked by a disaster, assume you will be flooded with traffic more than you expect your application to receive, assume your application will fail at one point or another. By incorporating these principles into the design and implementation of the application and
building mechanisms to handle the failures, a fault tolerant cloud application can be architectured [84].

enStratus’s chief technology officer (CTO) George Reese, offers some basic requirements for designing cloud applications for failure in [85].

- Deploy each application component across redundant cloud components and eliminate any common point of failure if possible.

- No assumption should be made about the underlying infrastructure of the cloud system. That is, a change in the underlying infrastructure should have minimal impact on the application uptime.

- Each application must be able to survive network outage among the nodes that support that component.

- Tools for facilitating automatic application responses to failures or other changes in the infrastructure, should be in place.

7.3.2 Determine Which Applications are Mission Critical

Cloud customers must work closely with cloud providers to determine the criticality or desired level of availability for each business applications moved to the cloud. Additionally, the maximum tolerable data loss (RPO) and maximum tolerable service disruption (RTO) should be determined. For instance, mission critical business applications require high level of availability with a state-of-art data back up and disaster recovery solution, since without access these applications, business operations are affected resulting in loss of revenue [86].

7.3.3 Multiple Cloud Providers

A simple way to mitigate risk of service failures is by using multiple cloud service providers. By deploying the same service across multiple cloud providers, the risk of service failures caused by power outage, network outage or other data center issues is prevented since the cloud providers would be operating independently. The downside to this approach is higher cost since the same capacity and capability is required across service providers. In the event that one of the providers shut down due to bankruptcy or other reasons, data is not locked-in and business continuity is not affected [87].
7.3.4 Cloud Monitoring

Unavailability or poor performance of cloud applications can result in loss of revenue for both cloud service providers and their customers. Proactively monitoring cloud applications and identifying glitches before end users experience any service lag makes sure the problems are fixed even before end users know they exist. If the problem originated from the cloud service provider, the customer can immediately notify them about the issue. Monitoring applications hosted on the cloud introduces a bit of complexity compared to the ones hosted within the traditional network perimeter of an organization. To simply put it, we are monitoring an application outside our local network, across the internet, within the cloud infrastructure of the cloud service provider. It could so happen that a network node is down along the traffic delivery chain. This is similar to the problem we experienced during the service availability test conducted in this thesis. We were getting false “down” reports for our cloud applications while the alerts were really as a result of network issues on the part of the monitoring service provider.

A simple approach to performing availability and integrity check on cloud storage services is by writing a small piece of data to the cloud storage service while keeping a copy, reading the data back and then comparing the read-back with the copy. This can be automated from servers in different locations.

A number of third parties monitoring service providers such as Amazon CloudWatch, LogicMonitor, Monitis, Gomez, TechOut, UptimeRobot and Zenoss can significantly save customers the time and cost of setting up in-house monitoring tool.

7.4 What End users of Cloud Services Should do to Improve Service Availability

There are a number of ways end users of cloud services can improve service availability.

7.4.1 Uninterrupted Internet Access

Cloud computing is a technology that uses Internet with the help of remote servers that host several data and applications for usage by the end users. Cloud computing gives the opportunity to various categories of users whether for personal or commercial to access their files at any computer without any installation using Internet access. This technology allows the usage of computing resources such as storage, memory, processing and bandwidth in a much more efficient manner. These files need high speed Internet connection
to access them and any form of low bandwidth will result in unavailability of that data at that point in time. Internet services are supplied by the Internet Service Provider (ISP) and to perform cloud efficiency those files require non-stop Internet connection.

7.4.2 Adaptability and Learning

The biggest issue for every user when trying to deal with cloud computing is on learning the application presented by the enterprise. This usually happens most especially when the application influences the user experience through online application. Every new application, users have to get familiar with it before it can utilize the online application which pose some kind of danger on the side of the users. Cloud computing is being adopted for example the business leaders whereby may be the specific application is used for online purchase and needed to be configured properly for its maximum efficiency and availability of all its services contents, attempt by the users not properly aware how it works will affect some part of application availability.

7.4.3 Feedback Report

End users should cultivate the habit of giving the feedback to their cloud services providers because this will enable the provider to know the recent threat, bugs and fix by sending them newest patches which increases application availability. This will enhance quick resolve automatically sometimes when the provider already aware of such threat in the cloud availability region.

7.4.4 Password Management

Several password management systems have been developed nowadays that will eliminate the headache of forgetting different passwords used by multiple applications in the cloud. For example McAfee Cloud Single Sign On system relieves the pain that end users have managing multiple passwords for cloud applications. This system allows the end user to enforce corporate standards for cloud application access and improves productivity, maintaining availability for IT and end users by eliminating password reset request.
8 Limitations, Conclusion and Future Work

This chapter begins by discussing the thesis limitations. It continues with a restating the aim of the thesis and the findings arrived at. It concludes by suggesting areas where the thesis can be further expanded upon.

8.1 Thesis Limitations

There were some limitations in the research. First, we conducted a test which entails deploying a web application to three different cloud computing platforms. The goal was to monitor the uptime of the applications for at least a month and to compare these metrics with the uptime guaranteed by each cloud platform. While each of the three cloud service provider offers monthly connectivity uptime percentage, it would be better if we were able to run the test for a longer time, ideally, for a year, since uptime is usually calculated on a yearly basis. Also, due to limited computational resources, the applications were monitored at 5 minute intervals, which means a service interruption could have occurred and recovered from within the 5 minutes without us finding out. Second, we conducted a survey among cloud service providers in the Nordic region, regarding availability issues in cloud computing. The population of the respondents were very low. In fact, only 13% of the recipients answered the survey which did not give us much findings for the research.

8.2 Conclusion

The goal of this thesis as previously in the aims and objectives was to investigate the threats to the availability of service in cloud computing and to provide some best practices that can be used to mitigate the threats. The motivation lies in the fact that availability is still a growing concern for organizations who already adopted cloud computing and one of the top reasons why other organizations are not adopting this technology.

Primary and secondary materials were used in the thesis. Primary materials being survey and testing. The survey entails using questionnaires to get the cloud service providers’ perspectives on availability issues in cloud computing. The testing entails a first-hand investigation of the service availability targets of three major cloud providers. Secondary materials derived from journals, articles, technical briefs, research papers, books and publications were used to formulate the initial findings in the theoretical part of the work. Sources such as news articles, security incident reports
and press releases were used to investigate the current threats facing
availability of cloud services.

As a result, out of more than 40 cloud service outages and failures
examined, 8 categories of threats were formulated. They include, in no
particular order: power outage, hardware failure, cyber-attack,
configuration error, software bug, human error, administrative or legal
dispute, and network dependency. It goes without saying that cloud service
providers can do more in terms of transparency about the cause of service
failures and effective communication with customers during service
disruptions.

A number of techniques centering around redundancy, that can be used
by cloud service providers to mitigate some of these threats, were proposed
in section section 7.3. Similarly, customers of cloud service providers do
not have to solely rely on providers to ensure constant availability, they can
employ a number of techniques including, but not limited to, designing their
applications to be fault tolerant and the use of multiple cloud providers.
In addition, techniques that can be used by end users of cloud services to
improve availability and ensure continuous access to their data in the cloud,
were provided in section 7.4.

Taken together, these results suggest that loss of availability or service
failures is a prominent threat in cloud computing, and no cloud service
provider is too big to fail. However, despite the challenges faced, many
organizations are increasingly adopting cloud computing to maximize
business returns.

8.3 Future Work

As mentioned in the limitations, the test conducted during the thesis could
be run for a longer period of time, preferably, a year, since uptime is mostly
calculated on a yearly basis. A robust monitoring tool can be built to monitor
the applications, instead of the third party monitoring services employed in
the test. In addition, the survey can be expanded to cloud service providers
around the world and not only the ones in the Nordic region.
References


<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AWS</td>
<td>Amazon Web Services</td>
</tr>
<tr>
<td>CDN</td>
<td>Content Delivery Network</td>
</tr>
<tr>
<td>CERT</td>
<td>Computer Emergency Response Team</td>
</tr>
<tr>
<td>CRM</td>
<td>Customer Relationship Management</td>
</tr>
<tr>
<td>CTO</td>
<td>Chief Technology Officer</td>
</tr>
<tr>
<td>DDoS</td>
<td>Distributed Denial of Service</td>
</tr>
<tr>
<td>DoS</td>
<td>Denial of Service</td>
</tr>
<tr>
<td>GAE</td>
<td>Google App Engine</td>
</tr>
<tr>
<td>HTTP</td>
<td>Hypertext Transfer Protocol</td>
</tr>
<tr>
<td>IaaS</td>
<td>Infrastructure as a service</td>
</tr>
<tr>
<td>IDC</td>
<td>International Data Corporation</td>
</tr>
<tr>
<td>IEC</td>
<td>International Electrotechnical Commission</td>
</tr>
<tr>
<td>IEEE</td>
<td>Institute of Electrical and Electronics Engineers</td>
</tr>
<tr>
<td>ISMS</td>
<td>Information Security Management Systems</td>
</tr>
<tr>
<td>ISO</td>
<td>International Organization for Standardization</td>
</tr>
<tr>
<td>ISP</td>
<td>Internet Service Provider</td>
</tr>
<tr>
<td>IT</td>
<td>Information Technology</td>
</tr>
<tr>
<td>IWQCR</td>
<td>International Working Group on Cloud Computing Resiliency</td>
</tr>
<tr>
<td>MTBF</td>
<td>Mean Time Between Failures</td>
</tr>
<tr>
<td>MTTR</td>
<td>Mean Time To Repair</td>
</tr>
<tr>
<td>NIST</td>
<td>National Institute of Standards and Technology</td>
</tr>
<tr>
<td>OWASP</td>
<td>Open Web Application Security Project</td>
</tr>
<tr>
<td>PaaS</td>
<td>Platform as a service</td>
</tr>
<tr>
<td>RAID</td>
<td>Redundant Array of Independent Disks</td>
</tr>
<tr>
<td>RAIS</td>
<td>Redundant Array of Inexpensive Servers</td>
</tr>
<tr>
<td>RPO</td>
<td>Recovery Point Objective</td>
</tr>
<tr>
<td>RTO</td>
<td>Recovery Time Objective</td>
</tr>
<tr>
<td>SaaS</td>
<td>Software as a service</td>
</tr>
<tr>
<td>SAF</td>
<td>Service Availability Forum</td>
</tr>
<tr>
<td>SLA</td>
<td>Service Level Agreement</td>
</tr>
<tr>
<td>TLS</td>
<td>Transport Layer Security</td>
</tr>
<tr>
<td>UPS</td>
<td>Uninterrupted Power Supplies</td>
</tr>
</tbody>
</table>