ABSTRACT

Security Analysis and Framework of Cloud Computing with

Parity-Based Partially Distributed File System

By Ali Asghary Karahroudy

July, 2011

Director of Thesis or Dissertation: Dr. M.H.N. Tabrizi

Major Department: Department of Computer Science

Abstract - *Cloud computing offers massive scalability, immediate availability, and low cost services as major benefits, but as with most new technologies, it introduces new risks and vulnerabilities too. Despite the fact that different cloud structures and services are expanding, the cloud computing penetration has not been as envisioned. Some specific concerns have stopped enterprises from completely joining the cloud. One of the major disadvantages of using cloud computing is its increased security risks. In this study I conduct an in depth analyses of the different aspects of security issues in cloud computing and propose a file distribution model as a possible solution to alleviate those security risks. It also shows the effectiveness of the new security model as compared with those currently being used. I present, a new file storage system with variable size chunks, distributed chunk addressing, decentralized file allocation tables, spread deciphering key, randomly selected file servers, and fault tolerant chunk system.*

Security Analysis and Framework of Cloud Computing with Parity-Based Partially Distributed File System

A Thesis/Dissertation

Presented to the Faculty of the faculty of Department of Computer Science

East Carolina University

In Partial Fulfillment of the Requirements for the Degree

Masters in Software Engineering

By

Ali Asghary Karahroudy

July 2011

Ali Asghary Karahroudy

Aliakarahroudy@Gmail.com

Security Analysis and Framework of Cloud Computing with

Parity-Based Partially Distributed File System

By

Ali Asghary Karahroudy

APPROVED BY:

DIRECTOR OF THESIS:____

M.H. Nassehzadeh Tabrizi, PhD

COMMITTEE MEMBER:____

Sergiy Vilkomir, PhD

COMMITTEE MEMBER:____

Junhua Ding, PhD

COMMITTEE MEMBER:____

Qin Ding, PhD

CHAIR OF THE DEPARTMENT OF COMPUTER SCIENCE:

____

Karl Abrahamson, PhD

DEAN OF THE GRADUATE SCHOOL:

____

Paul J. Gemperline, PhD

Dedicated to

My Parents,

For all their support and love

ACKNOWLEDGMENT

I owe my deepest gratitude to

Prof. *M.H.N. Tabrizi*

Without whom this thesis would not have been possible, many thanks to him for his help and never ending support.

TABLE OF CONTENTS

I. List of Figures

1. Chapter 1: Introduction……………………………………………..…………1

2. Chapter 2: Cloud Computing……………………………………….…………5

2.1.Cloud Computing Structure…………………………………………….………6

2.1.1. Software as a Service (SaaS)……………………………….………...7

2.1.2. Platform as a Service (PaaS)………………………….…………...8

2.1.3. Infrastructure as a Service (IaaS)……………………..………....9

2.1.4. More Services offered by Cloud Computing …….………..……....9

2.2. Cloud Providers…………………………………………….…………10

2.3. Cloud Usage Types………………………………………...…………..…10

2.3.1. Virtualized Desktop on Client……………………………..……..10

2.3.2. Browser and Cloud Application…………………………..………10

2.3.3. Cloud Choreography……………………………………..….….……11

2.3.4. Load Demand Handling………………………………………..……..11

2.3.5. Dynamic Resource Allocation……………………………..…………11

2.3.6. Hybrid IT Systems………………………….….……...………..…..12

2.3.7. Cloud users examples……………………….…………………..12

2.4.Advantages of cloud computing…………………………….……….….……12

2.5.Disadvantages of cloud computing……………………..………………..…14

2.6.Public and Private cloud……………………………..………………......15

2.7.Current Situation………………………………………..…………………16

3. Chapter 3: Cloud Computing Security..………………………….……………………19

3.1.Security requirements…………………………………………………….………21

3.2.Dynamic information security…………………………………………..…23

3.3.Network security…………………………………………………………23

3.4.Virtualization Security……………………………………………………24

3.5.Security risk prevention………………………………………………..26

3.6.Encryption scheme……………………………………………………..…27

3.7.Security practices…………………………………………………………27

4. Chapter 4: Security analysis in cloud computing……………………………....29

4.1.Traditional security……………………………………………..……..29

4.2.Availability…………………………………………………………….…30

4.3.Third party data control……………………………………………..…30

4.4.Standard analysis…………………………………………………………31

4.5.Identifying assets…………………………………………………………32

4.6.Identifying threads and countermeasures…………………...……………32

4.7.Increased security issues…………………………………………………33

4.8.Data storage systems…………………………………………...……35

4.9.Google File System / Hadoop File System……………………………....37

5. Chapter 5: Partially Distributed File System with Parity……………….……43

5.1.PDFSP Components……………………………………………………...45

5.2.PDFSP Process……………………………………………………...…46

5.3.PDFSP security risk model…………………………………………......48

5.4.PDFSP Security risk model analysis…………………………………..……51

6. Chapter 6: Conclusion……………………………………...………………....55

7. References……………………………………….……….…..……….…..56

8. Appendices……………………………………………………………60

8.1.Appendix A: Some major cloud providers…………………………………..60

8.2.Appendix B: Cloud users examples……………………………………….62

I. List of Figures

1. Figure 1. Current and future level use of cloud services in organizations survey….16

2. Figure 2. Level of company commitment to SaaS …………………………...……18

3. Figure 3. Reasons prohibiting companies from migration cloud……………...……20

4. Figure 4. Security system in VM layer…………………………………...…….27

5. Figure 5. File chunks……………………………………………………..……….43

6. Figure 6. Chunk header………………………………………………..……….44

7. Figure 7. Decryption Key Structure…………………..…………………..………...45

8. Figure 8. PDFSP Process………………………………………………..……..…46

9. Figure 9. Key Transmission…………………………………….…….…………..47

10. Figure 10. Risk factor vs. N_s……………………………………...….…………..51

11. Figure 11. Risk factor vs. (n+K)…………………………..………..…………52

12. Figure 12. Availability vs. K……………………………..……….………..53

CHAPTER1: INTRODUCTION

The term “cloud computing” describes the computation architecture taking the form of a cloud which is easily accessible by users on demand. As a metaphor for the Internet, the “cloud” is a familiar world. In a system where there is no regard for what a system is composed of, and it is simply studied as a black box of hardware and software, the term “cloud” an identification role, but as soon as it is combined with “computing,” everything changes and the meaning becomes blurred and more generic. Some vendors and researchers define cloud computing narrowly as an updated version of utility computing; basically virtual servers that are available on the Internet right now. Other analysts paint broader strokes, arguing that anything that may be consumed outside of the firewall is “in the cloud,” including conventional outsourcing [1].

Cloud computing first appeared in 2006, when Amazon’s Elastic Computing Cloud (EC2) was introduced to the world. In 2007, Dell released its version of cloud computing at the same time that IBM’s Blue Cloud was introduced. Others such as Google’s Mapreduce, and Microsoft’s Windows Azure followed suit one after another**.** According to estimates, the cloud computing market share **c**ould reach $42 billion by 2012 [2].

Currently cloud computing is in its early stages, as it started with storage services and has flourished into full-blown applications. It has transformed the IT world; currently, for the most part, IT users must plug into cloud based services individually**,** but cloud computing consolidators or wholesalers are already emerging to allow for a cloud network experience.

In today’s economic environment as enterprises try to balance out and optimize their IT budgets, cloud computing may play an effective role in reducing the IT operations and management costs. It frees up critical resources and budgeting for discretionary innovative projects for increasing capacity or adding capability on the fly. This comes in when an existing application is facing expanding usage demands either in the quantity of users or the requests that are made. Several companies employ an IT team in order to maintain and expand their resources as they increase and this forces substantial and sometimes unnecessary budget allocations.

Typically, Enterprises have a 80/20 split between regular ongoing IT operations cost, including hardware, software licensing costs, development costs, and data center maintenance, etc. versus new investments for solving critical business needs - critical for the survival of business in these challenging times [3]. This is where cloud computing comes in and encompasses any subscription-based or pay-per-use service that is provided in real time on Internet that replaces the need to allocate IT expenses.

Cloud computing offers massive scalability, immediate availability, and low cost services as major benefits, but as with most new technologies, it also carries with it inherent new risks and vulnerabilities too [4]. There are different cloud structures and services that are expanding, but the cloud computing penetration has not been as it was initially envisioned. Some specific concerns have halted enterprises from joining the cloud. The major one is security concerns. Security is playing a major role in companies' migration (or lack thereof) to cloud computing.

Moving data into the cloud offers great convenience where users are not required to be knowledgeable about storage capacity, storing techniques, hardware management, or data maintenance. Online storage services of the cloud are offering flexible and customizable services that make data storage and management much easier and more accessible.

In the process of migrating from traditional computing methods to cloud computing, availability and integrity of data as well as security and privacy issues are of great importance. Users are more skeptical of the cloud service concept when they perceive that they must relinquish full control over their data. Incidents such as the recent downtime of Amazon's S3 [5] are very clear examples of these concerns. From a user's perspective, Confidentiality, Integrity, and Availability (CIA) [6] concepts are what should be expected from the cloud provider and this will serve as a very important decision making gauge for choosing cloud services (or not).

It is almost impossible to apply the traditional security methods as they are normally implemented in traditional computing models with the cloud environment because of the deferred control of data from the customer to the cloud provider, and the unknown number of third parties between the real cloud service provider and the real customer. This imposes the need for new security and data distribution models that is compatible with the characteristics of cloud computing environment. In this thesis, a data security model will be introduced that fulfills the CIA requirements of data security.

To understand the dilemma more clearly and in more detail, the security concern will be analyzed and the taxonomy will be reviewed and a standard analysis based upon the ISO 27005 will be introduced. Then a prediction of future concerns as the cloud computing field develops will be presented. Finally, a data distribution model will be introduced to address the major concerns about data security. This thesis is organized as follow:

Chapter 1 provides an introduction to the thesis.

Chapter 2 provides a survey of cloud computing. As cloud computing is a relatively new technology, it is necessary to provide the reader with a background on the extensive components of this resource, such as the platforms, services, infrastructure, architectural points, applications, providers, scalability, availability, and advantages. While the scope of the chapter may appear to be broad, it provides the reader with a vital introduction and a bridge to subsequent chapters, and it would greatly detract from the thesis without it.

Chapter 3 discusses cloud computing security by detailing the complications and challenges associated with cloud computing, and the types of security protocols that are integral and required on both the “software” and “hardware” sides. It discusses information security, network security, virtualization security, and security risk prevention, and provides an analysis of all of the components of security in the cloud. This security analysis is continued in Chapter 4.

Chapter 4 continues the analysis of security in cloud computing, and details a standard analysis of Confidentiality, Integrity, Availability, Assets, Threats and Countermeasures, and concludes with the various types of data storage and file systems and an introduction to Chunk Servers.

Chapter 5 provides with the introduction to and explanation of the Partially Distributed Parity-Based System, including the components, the process, the security risk model, and the security risk model analysis.

Chapter 6 provides a conclusion and summary of cloud computing and the Partially Distributed Parity-Based System.

CHAPTER 2: CLOUD COMPUTING

It is a very common mistake to confuse cloud computing with what we refer it to the cloud or traditional web hosting. In a traditional cloud environment, the company to whom the consumers are paying their monthly fees has a bunch of server boxes (virtual or actual). Each of these has server software and a framework. When a website space is rented from service providers, they create a virtual directory application for the customer. So, on their system, the customer’s application might be at server12/customer1234, and they also provide a DNS mapping so that www.thiswebsite.com points to that location on their system and the end users never know about how this plumbing works.

The key architectural points to consider here are that there is a single physical point in which the customer application resides; there is a single database management instance, also hosted in a single physical location, and a single database within that database instance that contains the customer’s data. It is a classic example of the client/server model, with the server parts being hosted by the web host company and the client parts being browsers and other web consumers.

However, cloud computing is different from the traditional concept of the cloud and by the very nature in which it is distributed. When a user develops a web application in the cloud computing environment IDE, there might be five different instances of the application running in five different data centers throughout the world. Also, this application could be pulling data from partitioned storage in three different data centers throughout the world. There may be eight different instances of the "engine" running in eight different data centers, all working and processing the information that the user’s application is taking in and fanning out.

Another common mistake is to confuse cloud computing with grid computing. Cloud computing environments support grid computing by providing physical and virtual servers on which the grid applications can run. Cloud computing is different from grid computing because grid computing involves dividing a large task into many smaller tasks that run parallel on separate servers. Grids require many computers, typically in the thousands, and commonly use servers, desktops, and laptops. Clouds also support non grid environments, such as a three-tier web architecture running standard or Web 2.0 applications [7]. A cloud is more than a collection of computer resources because a cloud provides a mechanism to manage those resources. Management includes provisioning, change requests, reimaging, work load rebalancing, deprovisioning, and monitoring.

Writing a scalable, robust, reliable application in cloud computing involves being aware of the architecture of "the cloud," but developers do not need to worry about things such as how to implement hot failover. They do not have to worry about things such as what happens to the web service in case of mishap. They are able to just go to the admin site and bump up the number of running instances of the web role and maybe the worker role as well [8].

2.1 Cloud Computing Structure

One of the most important parts of cloud computing technique is the advent of cloud platforms. As its name suggests, this type of platform allows developers to write applications that run in the cloud, or use services provided by the cloud, or both. It is also known as *on-demand platform* and *platform as a service (PaaS)*. The difference between how application platforms are used today and cloud platforms is that when a development team creates an on-premises application (i.e. one that will run within an organization), it only needs to write a source-code in some type of programming language.

If the creators of every on-premises application first had to build all of the supporting software that is, right from the operating systems to the assembler who decodes the program, then we would have fewer applications in existence today. Similarly, if every development team that wishes to create a cloud application must first build its own cloud platform, we would not see many cloud applications either.

Fortunately, there are several cloud platform technologies available today [7] allowing entire businesses and thousands of employees to run their computing needs as online rented tools. All of the processing and file saving will be performed in the cloud, and the users will plug into that cloud every day to do their computing work [9]. Cloud computing has three major premises:

- Software as a Service
- Platform as a Service
- Infrastructure as a Service

2.1.1 Software as a Service

Software as a Service (SaaS) is software distributions model in which applications are hosted by a vendor or service provider and are made available to customers over a network, typically the Internet.

SaaS is becoming an increasingly prevalent service delivery model as underlying technologies that support web services and Service Oriented Architecture (SOA) mature, and new developmental approaches, such as Ajax become popular. Meanwhile, broadband service has become more widely available to support user access from more locations around the world.

SaaS is closely related to the Application Service Provider (ASP) and the on-demand software delivery models. IDC identifies two slightly different delivery models for SaaS [10].

The hosted Application Management (AM) model is similar to ASP: a provider hosts commercially available software for customers and delivers it on the Web.

Using the software on-demand model, the provider provides customers network-based access to a single copy of an application created specifically for SaaS distribution. Benefits of the SaaS model include:

- Easier administration
- Automatic updates and patch management
- Compatibility
- Easier collaboration
- Global accessibility

2.1.2 Platform as a Service

Platform as a Service (PaaS) is a way to rent hardware, operating systems, storage, and network capacity on the Internet. The service delivery model allows the customer the ability to rent virtualized servers and associated services for running existing applications or developing and testing new ones. PaaS is an outgrowth of SaaS, and it is software distribution model in which hosted software applications are made available to customers on the Internet.

PaaS has several advantages for developers where operating system features may be changed and upgraded frequently. Geographically distributed development teams may work together on software development projects. Services may be obtained from diverse sources that cross international boundaries. Initial and ongoing costs may be reduced through the use of infrastructure services from a single vendor rather than maintaining multiple hardware facilities that often perform duplicate functions or that suffer from incompatibility problems. Overall expenses may also be minimized by unifying development efforts [10].

On the downside, PaaS involves some risk of "lock-in" if offerings require proprietary service interfaces or development languages. Another potential pitfall is that the flexibility of offerings may not meet the needs of some users whose requirements rapidly evolve.

2.1.3 Infrastructure as a Service

Infrastructure as a Service (IaaS) sometimes is referred to as hardware as a service. It is a provision model in which an organization outsources the equipment used to support operations, including storage, hardware, servers, and networking components. The service provider owns the equipment and is responsible for housing, running and maintenance. The client typically pays on a per-use basis. The characteristics and components of IaaS include:

- Utility computing service and billing model.
- Automation of administrative tasks.
- Dynamic scaling.
- Desktop virtualization**.**
- Policy-based services.
- Internet connectivity.

2.1.4 More Services Offered By Cloud Computing

There are also some other services offered by cloud computing that include**:**

- Data as a Service (DaaS): The customer's queries against the provider's data base.
- Identity and Policy Management as a Service (IPMaaS): The provider manages identity and/or access control policy for customer
- Network as a Service (NaaS): The provider offers virtualized networks like a VPN.

2.2 Cloud Providers

Based on the structure and concepts described above and specifications on cloud providers in the market; some major cloud computing providers are listed in Appendix A.

2.3 Cloud Usage Types

Cloud computing applications are constructed from program logic and across-the-network calls to cloud services. According to [11], typical users of cloud computing services may be categorized as follow:

2.3.1 Virtualized Desktop on Client

This type of user could employ software such as Virtual Bridges VERDE server to run desktop applications on a powerful server, but have the screen output delivered down to a local device such as a netbook, laptop, or desktop machine. While many people speak of virtualized Linux desktops today, there remains a future vision where many organizations run native local Linux desktops and then virtualize down from the cloud. Windows desktop would remain only for light and occasional use of applications that have not yet been ported or replaced [11].

2.3.2 Browser and Cloud Application

This is a very broad view of how a user might use the "cloud." Here, the user would have a sense that instead of running a local application, software much like a word processor or CRM front-end is used in a browser such as Firefox**.** These cloud-based applications are helping to reduce users' dependencies on working on any particular operating system, and therefore allowing more and more use of Linux and Mac/OS X in businesses and organizations. On the other hand, traditional desktops may be extended to use the cloud for remote storage [11].

2.3.3 Cloud Choreography

The notion of choreography is borrowed from web services or Service Oriented Architecture (SOA). The idea is that new applications are constructed from program logic and across-the-network calls into cloud services. The excitement will come when more than one cloud is used, and therefore further emphasizes the need for open standards and cloud interoperability. Security issues always will be important, but privacy issues strongly enter into this scenario because of the possibility of improperly sharing information across services and clouds. This case most clearly shows where Software as a Service (SaaS) might be subsumed into the general notion of cloud computing [11].

2.3.4 Load Demand Handling

Service providers desire the ability to possess the right level of software and hardware resources to provide an acceptable quality of service to their customers. Cloud computing may be able to deliver this by allowing a service provider to purchase and configure datacenter resources for average use, and then to employ processors or storage from the cloud in order to handle spikes [11].

2.3.5 Dynamic Resource Allocation

While a software developer might spend a substantial amount of time thinking about and working in an integrated development environment where the need for computer resources is small, other activities such as compiling, linking, and testing may be very computer resource intensive. In those cases, a private or public cloud could be used so that the local capital expense for servers may be minimized [11].

It is very important to observe and measure how and when developers use computing resources before contracting cloud services. For example, does an entire workgroup of developers require the resources simultaneously, or do the individuals need them fairly randomly? In international efforts such as computer animations, will one shift of developers use resources that are no longer needed by others in a different time zone?

2.3.6 Hybrid IT System

This is the lowest “closest to the metal” level of user who uses clouds but who does not build them. Someone else configures the datacenter but it is this system administrator’s job to decide how to best deploy applications onto either traditional dedicated servers or shared cloud servers. This administrator would need to understand the resource needs of the applications, as well as the security parameters. As to the latter, may a given application be run on the same physical node as other software, or must it be isolated? How do corporate instructions about data security affect the decisions about how to deploy applications or to place storage? In these instances, the cloud becomes an additional IT option with its own characteristics that must be managed alongside traditional IT deployments.

2.3.7 Cloud Users Examples

Some examples of enterprises that deliver services based on cloud are listed in appendix B.

2.4 Advantages of Cloud Computing

Despite its possible security and privacy risks, cloud computing has benefits that the public sector and government IT organizations want to take advantage of, and in brief, they are as follows:

Location and device independence: Users are not tied to any device or location to access the service. All they need is a simple Internet connection and a web browser.

Reduced cost: Cloud technology is paid incrementally based upon, consumption, this means saving the organization money.

Increased storage: Organizations may store more data in the cloud than on a private computer or network system.

Highly automated: Cloud computing services are highly automated. This is based upon all efforts for maintenance and related issues that are made by the cloud service provider [12].

Flexibility and scalability: A change of resources in any way is possible and easy to implement mush faster than in past computing methods. IT departments that anticipate an enormous increase in user load does not need to scramble to secure additional hardware and software with cloud computing. Instead, an organization may add and subtract capacity as dictated by its network load. Better still, because cloud-computing follows a utility model in which service costs are based upon consumption, companies will only pay for what they use [13].

IT focus shift: No longer having to worry about constant server updates and other computing issues, IT will be free to concentrate on innovation [14].

Utilization and efficiency improvement: Based upon the cloud service provider, efficiency will be improved greatly, much more than it could possibly for a local IT team.

Easy implementation: Without the need to purchase hardware, software licenses or implementation services, a company may get its cloud computing arrangement off the ground in record time and for a fraction of the cost of an on-premise solution.

Skilled practitioners: When a particular technology becomes popular, it is not uncommon for a whole slew of vendors to jump on the bandwagon. In the case of cloud computing, however, vendors have typically have been reputable enough to offer customers reliable service and large enough to deliver huge datacenters with endless amounts of storage and computing capacity. These vendors include industry stalwarts such as Microsoft, Google, IBM, Yahoo, and Amazon.

Quality of service: Network outages may send an IT department scrambling for answers. Yet, in the case of cloud computing, it is up to a company's selected vendor to offer 24/7 customer support and an immediate response to emergency situations. That is not to suggest that outages do not occur. In February 2008, Amazon's S3 cloud-computing service experienced a brief outage that affected a number of companies. Fortunately, service was restored within three hours [12].

Greener technology: A typical data centre consumes up to 100 times more power than an equivalent sized office building. The carbon footprint of a typical data centre is therefore a significant concern for many organizations [15] that could be improved by adopting the cloud technology.

2.5 Disadvantages of Cloud Computing

Just as every other existing phenomena, cloud computing has its own disadvantages that usually vary from provider to provider and are very much dependent upon the service provider policies. The disadvantages are summarized as:

Customer control: The customer does not have control over the cloud provider's server, software, and security processes.

Data security: The customer's data is under the provider's control, so this requires strong trust between two parties, something that is usually very difficult to obtain.

Anchoring problem: It is almost impossible to migrate massive amounts of data from the service provider to any other party in the event of any need to change the provider, so the customers are almost anchored to their initial provider.

Hidden costs: In some cases, hidden costs such as additional cost of data transfer are applied.

Network dependability: If the network crashes, the customer will be unable to use the services until the network is restored and it is backed up.

Legacy compatibility: Some applications or hardware might require having a hard drive attached to the computer; these might be hard to get working properly with the hard drive on a remote server**.**

Security: In nearly every survey done about cloud computing the top reason given for not adopting it is a concern over security.

2.6 Public and Private Cloud

In a typical cloud computing scenario, organizations run their applications from a data centre provided by a third-party – the cloud provider. The provider is responsible for providing the infrastructure, servers, storage**,** and networking necessary to ensure the availability and scalability of the applications. This is what is referred to as a public cloud. There is also a private cloud. A private cloud is a proprietary computing architecture, owned or leased by a single organization that provides hosted services behind a firewall to "customers" within the organization. Some commentators regard the term "private cloud" as an oxymoron [15]. They say that the word "cloud" implies an infrastructure running on the Internet, not one hidden behind a corporate firewall. There is, however, a larger body of opinion suggesting that private clouds will be the route chosen by many large enterprises and that there will be substantial investments in this area. Already**,** vendors are lining up to release products that will enable enterprises to more easily offer internal cloud services.

2.7 Current Situation

Currently, startup companies and small businesses are showing the highest interest in cloud computing services. They are able to use the cloud for almost everything they need. Services such as software, infrastructure, collaboration services, and an online presence are what lead them to be interested. The situation is different with mid-size enterprises, they are able to use the cloud but some of the cloud services are still unclear to them. These companies usually benefit from compute cycles for R&D, online collaborations, partner integration, social networking, and new business tolls. Finally, large enterprises are more likely to have hybrid models where they maintain some of their computing activities in house, usually because of legal and risk management reasons.

A study from IDC [16] shows that in all application fields, there appears to be a tendency toward leaning to cloud computing technologies over the next three years, as shown in Figure 1 below:

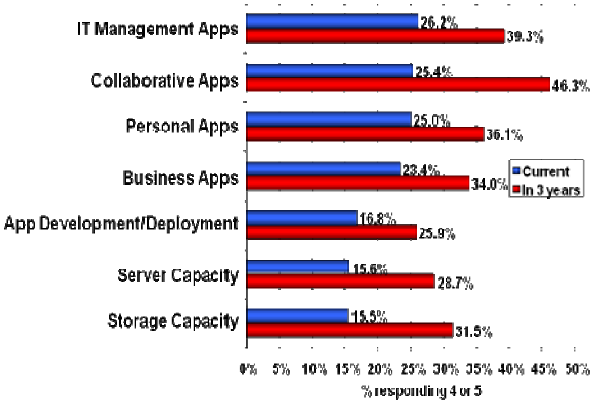

Figure 1: Current and future level use of cloud services in organizations survey

Source: IDC Enterprise Panel, August 2008 n=244

Also, considering the level of company commitment to SaaS applications as a percent of overall applications they use (IDC 2009) shows that only 1% of the related companies' state that SaaS is definitely not part of their application strategy at the time. This will be the time when 18.7% of the companies are 100% committed to the SaaS while the remaining companies are struggling to migrate from the old fashioned application system to the cloud computing environment.

Technology experts and stakeholders say they expect to 'conduct most of their business in the cloud' in 2020, working mostly with cyberspace-based applications accessed through networked devices. This will substantially advance mobile connectivity through/with smart phones and other Internet appliances. Several surmise/suggest that there will be a cloud-desktop hybrid; Still, cloud computing has many difficult hurdles to overcome, including concerns tied to the availability of the broadband spectrum, the ability of diverse systems to work together, security, privacy, and the quality of service.

Among the most current popular cloud services are social networking sites (the 500 million users of Facebook), webmail services such as Hotmail and Yahoo mail, micro blogging and blogging services such as Twitter and WordPress, video-sharing sites such as YouTube, picture-sharing sites such as Flicker, document and applications sites such as Google Docs, social-bookmarking sites such as Delicious, business sites such as eBay, and ranking, rating, and commenting sites such as Yelp and Trip Advisor [17]. It is nearly a consensus among more than 72% of survived parties that:

"By 2020, most people won't do their work with software running on a general-purpose PC. Instead, they will work in Internet-based applications such as Google Docs, and in applications run from smart phones. Aspiring application developers will develop for smart phone vendors and companies that provide Internet-based applications, because most innovative work will be done in that domain, instead of designing applications that run on a PC operating system." Figure 2 shows the real level of tendency to use cloud computing technology.

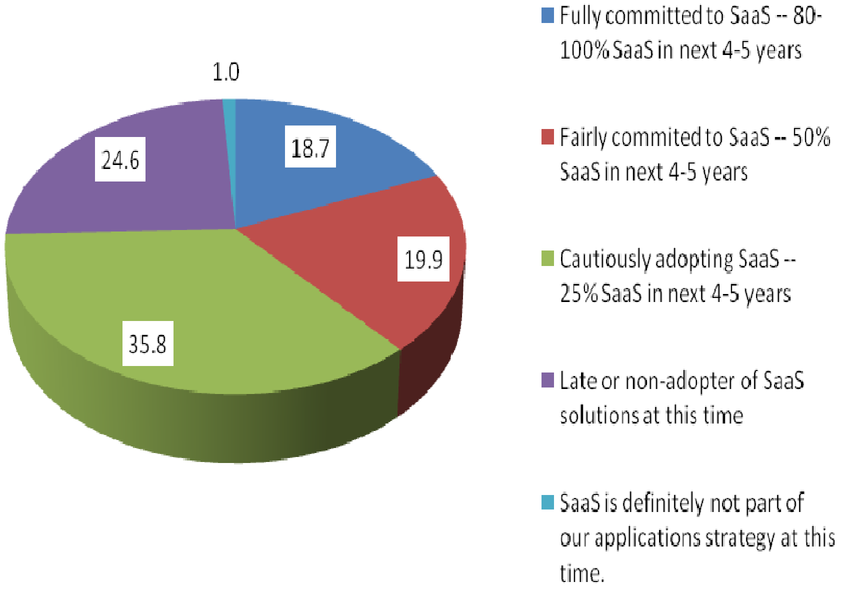

Figure 2: Level of Company Commitment to SaaS Applications as a Percent of Overall Applications Use

Source: IDC SaaS Adoption Survey, November 2008. n= 200

CHAPTER 3: CLOUD COMPUTING SEURITY

Surveys of senior-level IT managers indicate [18], security is consistently one of the top five concerns, along with, specifically, the security related to the available technology of the moment. Security concerns arise because both the customer data and program reside within the providers premises. In nearly every survey done about cloud computing the primary reason provided for not adopting is security concerns [15].

Putting business-critical data in the hands of an external provider still petrifies of most managers. Only by relinquishing some control over the data will companies then capture the cost economies that are available after joining the cloud computing technology. A company must determine when the trade-off is worthwhile. In deciding on the trade-off some of the questions to consider are:

- What happens if the data stored or processed on a cloud machine becomes compromised?
- Will the customer be informed of that?
- If the customer does not know, how will they notify their constituents, especially when data breach notification laws are in place?
- How will the customer know to improve their security?

The truth of the matter is that holding the data in the cloud is not really any less secure than leaving it on internal servers connected to the Internet. The recent case in the UK [15], about a hacker who hacked his way into the US Government network shows that seemingly secure networks are just as likely to be breached. Companies need to be realistic about the level of security they may achieve inside of their own business, and how that might compare to a cloud provider. It is well known that more than 70% of intellectual property breaches are a result of attacks made from within the organization.

Despite this, security will be raised as a concern regarding cloud computing for many years to come. There is still much work to be done before more formalized standards are set in place. Organizations such as the Cloud Security Alliance are at the forefront of addressing these issues. In the same way that some banks have hesitated longer than others in offering Internet banking facilities so shall it be with cloud computing. Some organizations may evaluate the risks and may adopt cloud computing quickly, while other more conservative organizations may be more apt to observe from the "sidelines" and watch the developments unfold [15]. In addition, a recent study by IDC [19], shows that about 88.5% of the customers that are likely to avoid using cloud computing cite security as the main reason for this denial (see Figure 3).

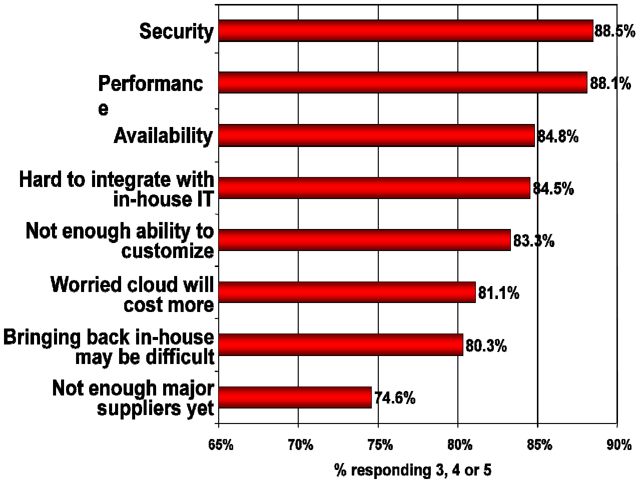

Figure 3: Comparison between reasons prohibiting companies from migration to cloud

Source: IDC SaaS Adoption Survey, November 2008. n= 200

Security is usually defined as saving data and program from danger and vulnerability. Dangers that threaten the data are:

- Disruption of services.
- Theft of information.
- Loss of privacy.
- Damage of information.

And the most crucial vulnerabilities listed would be:

- Hostile program.
- Hostile people giving instructions to good programs.
- **“**Bad guys**”** corrupting or eavesdropping on communications

3.1 Security Requirements

Security is needed at the different levels that include:

- Server access security**.**
- Internet access security**.**
- Database access security**.**
- Data privacy security**.**
- Program access security**.**

Applying security protocols includes both the “software side” security and the “hardware side” security. A good cloud computing provider must have secure enough policies in place to keep the data safe from the dangers and vulnerabilities stated in the previous section. Some of the important security requirements are:

Confidentiality: Ensuring that information is not disclosed to any unauthorized parties.

Integrity: Ensuring that information held in a system, is a proper representation of the information intended and that it has not been modified by an unauthorized person.

Availability: Ensuring that information processing resources are not made unavailable by malicious action.

Non-repudiation: Ensuring that agreements made electronically may be proven to have really transpired.

Physical security: On the "hardware side" of security there are several well defined protocols in the industry, such as the professional security staff utilizing video surveillance, state of the art intrusion detection systems, and other electronic means for guarding a datacenter. Furthermore, when an employee no longer has a legitimate business purpose for accessing the datacenter, that employee's privileges for accessing the datacenter should be immediately revoked, Physical security protocols should be applied to all datacenters and backup centers and wherever user data is stored or used.

Data sanitization: Sanitization is the process of removing sensitive information from a storage device. What data sanitization practices does the cloud computing service provider propose for implementing redundant and obsolete data storage devices when and if these devices are retired or discontinued.

The "software side" of security has guided us to a deeper and newer era. For more than a century, physical security has existed and has been developed day after day and has continuously evolved, but on the "software security side," science is still young and evolving. This makes it a challenge for cloud computing developers. A fresh field of research and study which should answer the most critical questions:

- What is data security at the physical layer?
- What is data security at the network layer?
- What about investigation support?
- How safe is data from natural disaster?

- How trustworthy is the encryption scheme of service provider?
- How secure is the code?
- And etc.

3.2 Dynamic Information Security

Information security, as static information residing on hard disks at datacenters or backup centers should be fulfilled by physical security protocols, but what about data that is being transferred from one host to another? Security related to the information exchanged between different hosts or between hosts and users is provided by transfer protocols and middleware. These are security issues pertaining to secure communication, authentication, and issues concerning single sign on and delegation. Secure communication issues include those security concerns that arise during the communication between two entities, and these include confidentiality and integrity issues. Confidentiality indicates that all data sent by users should be accessible only to "legitimate" receivers, and integrity indicates that all data received should only be modified by "legitimate" senders.

3.3 Network Security

Inherited from network security issues, cloud computing also suffers from some very well known security defects such as:

QOS violation: Resource hacking could delay or drop wanted packets through congestion. Spamming and a handful of techniques are used by hackers to stop or decrease the service quality.

Denial of service: Where server and networks are brought down by a huge amount of network traffic and users are denied access to certain Internet based services;

Spamming and a handful of techniques are used by hackers to stop service or decrease the service quality.

Sniffing: Also recognized as man in the middle attack happens when a hacker sniffs the packets and retrieves user data from them. SSL is used to prevent Sniffing.

Spoofing: Is when a hacker disguises as a arbitrary user IP address and creates packets that have someone else's IP address. One of the solutions to spoofing is force infrastructure not to permit an instance to send traffic with a source IP or MAC address other than its own.

Port scanning: Some specific ports with low profile security configurations that allow traffic from any source (very common configuration on most of the machines!) will be vulnerable to a port scan**.** When port scanning occurs, it should be detected by the network and stopped immediately.

ARP cache attack: attacker plugs into the network and sniffs the network traffic while he is injecting ARP messages into the network. ARP finds the MAC address associated with its particular IP address and then hacker sniffs the IP address.

3.4 Virtualization Security

The virtualization layer (or hypervisor), is effectively another operating system in the data center. Hypervisors tend to carry a much smaller footprint than a traditional operating system with a correspondingly lower potential for security holes. Plus, no hypervisor will be found surfing the Internet and downloading code or used as a station by any means.

Yet, at the same time, it is still a relatively immature product, and vulnerabilities are still repeatedly discovered. These vulnerabilities are usually quickly rectified, but should be monitored and tracked [20]. The maturity of hypervisor technology also shows in its vetting and certificating infrastructure. It is theoretically possible for hackers to attack the hypervisor layer specifically, or to take over a VM and use it to attack other VMs, according to Chris Steffen, principal technical architect at Kroll Factual Data, a credit-reporting and financial-information services agency in Loveland, Colorado. However, this has never happened "in the wild," so the threat remains theoretical for now, until that “someday” when it is detected in the real world.

No matter what virtualization operating system is chosen (VMware ESX, Virtual IRON, or others), a new operating system is introduced on the network and that is always a cause for security concern. There could be security holes that are in need of patching, and the possibility exists for guest to guest attacks; because virtual guest systems tend to move around with high availability or load balancing, these virtual guests may be difficult to keep track of, causing them to be more difficult to secure. Also, new virtual guest operating systems are so quick and easy to add, it may also be difficult to keep track of these new systems that are brought online, and therefore, more difficult to secure.

Another aspect of a virtualization security risk is the communication between virtual machines. Virtual machines have to communicate and share data with each other, and as noted by Ruykhaver [21], if these communications aren’t monitored or controlled they are ripe for attack**.**

As virtualization becomes more and more popular in market, it also will become more and more popular as a target for malicious attacks. As virtualization administrators, there is a necessity to ensure that these virtualized systems are as secure as or even more secure than our physical systems. Plus, there needs to be a demand for more and more security features from the manufactures of the hypervisors and virtualization management interfaces. In summary, virtualization is truly invaluable to us all; it is here to stay. Similar to wireless LANs, virtualization is a young technology and it needs more maturity in the area of security [21].

3.5 Security Risk Prevention

A Virtual Machine Monitor (VMM) is a host program that allows a single computer to support multiple, identical execution environments. According to [22]**,** a good VMM should support the following properties:

Isolation: Software running in a virtual machine may not access or modify the software running in the VMM or in a separate VM.

Inspection: The VMM has access to all of the state of a virtual machine: The CPU state (e.g. registers), all memory, and all I/O device state such as the contents of storage devices and register the states of I/O controllers, so that VMM may monitor VM.

Interposition: Fundamentally, VMMs need to interpose on certain virtual machine operations (e.g. executing privileged instructions). An example of this would be if the code running in the VM attempts to modify a given register. Therefore, a cloud computing system needs a system to help control the following items:

- Memory and CPU
- Networking
- Process execution control
- Storage

This system could be software (such as an antivirus) or hardware or combination of the two such as shown in Figure 4.

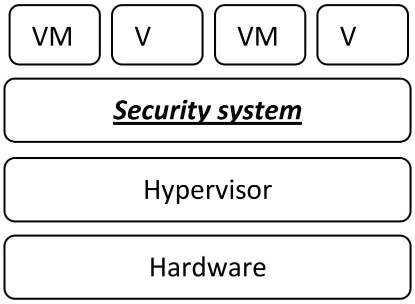

Figure 4: Security system in VM Layer

3.6 Encryption Scheme

A critical aspect of a security system is that every security system uses an encryption schema; the more secure this schema is the more accurate and safe a system may be in order to serve its purpose. In this sense, the security system should answer the following:

- Is it possible to encrypt all of the customers' data?
- What algorithms should be used?
- Who creates and holds the algorithm?
- Who holds the ciphering key?

Answering these questions seems so easy at first glance but they actually pose a very serious challenge. Encryption accidents may cause data to be corrupted and totally unusable; they may complicate accessibility and availability of data. In some cases, encryption may prevent applications from their normal functionality. The cloud computing company or the service provider is responsible for designing and testing an encryption schema and putting that in a virtual machine in order to make data safe.

3.7 Security Practices

Because of the cloud computing structure, investigating and searching for malicious activities or virus attacks is a complicated and difficult action, logging and data for multiple customers may be co-located and also may be geographically spread across an ever-changing set of hosts and data centers. A known solution to this issue is to secure a contractual commitment in order to support specific forms of investigation, along with evidence that the vendor has already successfully supported such activities [22], but even with having this legal advantage, technical difficulties still remain and hinder an effective, cost efficient investigation process.

Therefore, it is extremely difficult for the customer to actually verify the currently implemented security practices and initiatives of a cloud computing service provider because the customer generally has no access to the provider's facility which can be comprised of multiple facilities spread around the globe.

CHAPTER 4: SECURITY ANALYSIS IN CLOUD COMPUTING

The Cloud *S*ecurity Alliance's initial report [23] contains a different sort of taxonomy based on 15 different security domains and the processes that need to be followed in an overall cloud deployment. It categorizes the security concerns as:

- Traditional security
- Availability
- Third-Party data control

4.1 Traditional Security

Traditional security concerns, involve attacks on computers and networks that will be made easier when the resources are moved to the cloud. In a multi-tenant architecture, potential vulnerabilities in hypervisor or VM are the most challenging concerns. Some other weak points such as SQL-Injection or cross-site scripting vulnerability recently have been discovered in Google docs [24], and there could be other concerns in this category. Another incident in SalesForce.com [25], displays how fishers and other social engineers have a new attack vector after the cloud computing technology is introduced.

A unique security concern occurs when the cloud attacks the machine connected to it. The cloud user must protect the infrastructure used to connect and interact with the cloud, despite the fact that the cloud is often located outside of the firewall in many cases. An example of such expanded network attack surface is shown [26].

Authentication, authorization, and possible forensics in the cloud are other concerns in this category. Inspecting for forensics is not that easy when the inspector is dealing with cloud architecture. The blog posting the CLOIDIFIN [27], Project summarizes this difficulty.

4.2 Availability

Availability and maximum up-time concerns are centered on critical applications and the availability of data for users. Some outage incidents including Gmail [28], Amazon S3 availability issue [29], and FlexiScale [30], and their consequences reiterate the importance of this concern.

Despite the fact that cloud providers claim that their server up-time compares well with the availability of the cloud user's own data center, up-time is still the primary concern with respect to availability [31]. While cloud services have the appearance of providing more availability, this is not perhaps the case, because there are more incidents of failure and attack. Yet another concern [31] is the ability to assure enterprises that the cloud provider is faithfully running a hosted application and providing valid results.

4.3 Third-Party Data Control

Cloud computing means refers to data that is held by a third party (the cloud provider), which prompt legitimate, ample concerns for customers and a potential lack of control and transparency. Also, legal issues and required customer consent forms are propelling this concern further.

Some of the questions that are being asked are:

- What happens if customer's data is lost?
- Is the customer able to prove the deletion of the data?
- Is the data loss traceable?
- How is it possible to audit the cloud provider without violating other aspects of security?
- What are the contractual obligations?

Cloud provider espionage [32], is yet another concern; worries about the loss or theft of company proprietary information by the cloud provider. A CNN article [33] notes: For Shoukry Tiab, the vice president of IT at Jenny Craig, who uses Postini and Google Maps, the primary concern is security and confidentiality. He states, "Am I nervous to host corporate information on someone else's server? Yes, even if it's Google."

Another possible concern is that the cloud provider may be a cloud customer yet another provider itself and is cascading the service, which complicates the trust issue even more.

4.4 Standard Analysis

The term “information security” is usually used in an ambiguous way and/or sometimes incorrectly. Information security refers to the protection of any information system (Assets) from any unauthorized access. Main fields, sometimes called principles of information security, are:

- Confidentiality; prevent unauthorized disclosure
- Integrity; reserve information integrity
- Availability; ensure information availability when required

The above three items are referred to as CIA. Based upon the ISO27005 standard [34], identifying assets is the first step in security analysis process, and it is essential to know which assets are the ones that require protection and what properties of these assets must be maintained. The next step will identify threats, what types of attacks may be mounted, and what other threats exist. This will also lead to the identification of countermeasures that show ways to counter those attacks. Finally it should be noted that there is no established organizational context or policies and the study should be performed by an appropriate organization; through an independent analysis designed based on specific subject.

4.5 Identifying Assets

In a cloud computing enterprise the following assets could be recognized:

Customer Data; Data may be of any type or format, usually with no limitations. The customer transfers and maintains his own data on enterprise servers. Employee tables and payroll information, etc. are some prime examples.

Customer Application; Any Application hosted by an enterprise that is customer owned is considered an asset. Applications may come of any form, any allowed programming language and processes. There may be some limitations applied by enterprise.

Client Computing Devices; Connected devices to the cloud by users in order to benefit cloud services are also considered assets. A device which becomes inaccessible means the cloud service is inaccessible.

For all the three assets, all three CIA principles should be maintained.

4.6 Identifying Threats and Countermeasures

Any activity that endangers the CIA principles is considered a threat, and therefore a threat may present itself in different ways. A failure in provider security, a customer on customer attack, availability and reliability issues, legal and regulatory problems, a broken firewall, a defective security perimeter, and even the integration of provider and customer security systems and many more. Avery accurate and deep study over the threats is a must in security analysis of cloud computing (or any other field).

Based upon the identified threats, countermeasures such as verify and monitor cloud provider security by third party inspectors, hypervisors for compute separation, MPLS, VPNs, VLANs, firewalls, ant viruses, cryptography and key management, application layer separation, planned downtime, sharing resources, restricting geography, changing perimeter models, integration monitoring, and a myriad of other countermeasures that may be implemented.

It is important for security providers to engage in a full risk management process for each detailed case. For small and medium organizations, cloud security may be a huge improvement and the cost savings may be huge for large organizations that already have large secure data centers, where this cost effectiveness is not that visible. These companies generally look for elastic and secure services.

4.7 Increased Security Issues

Some increased security issues are based in development and therefore an extension of cloud computing. A fraction of these issues already exist to some extent and other issues may arise after the maturation and more widespread adoption of cloud computing technology.

> *Extra resources for attackers*: As cloud computing resources become more available**,** massive computational power, access to huge databases, information, data mining tools, and directories will assist in the attackers' effectiveness.
>
> *Privacy issues:* Because all of the data will be housed on the provider's premises, access to that data even through the normal search engines, such as Google search will be easily available. It is obvious that anonymizing data is a difficult problem to be solved. An example of indirect data-mining that might be performed by a cloud provider is to note transactional and relationship information (see World Privacy Forum Report) [35]. For example, the sharing of information by two companies may signal that a merger is under consideration.

Cost effectiveness of availability: Availability also needs to be considered in the context of an adversary whose goals are simply to sabotage any activities. Increasingly, such adversaries are becoming more realistic as political conflict is spilling over onto the web, and as the recent cyber attacks on Lithuania confirm [36]. The damages are not only related to losses in productivity, but extend to losses as a result of the lack of trust in the infrastructure, and the potentially costly backup measures. The cloud computing model encourages single points of failure. It is therefore important to develop methods for sustained availability (in the context of attack), and for recovery from attack. The latter could operate on the basis of minimization of losses, required service levels, or similar measures [34].

Authorization issues: As the adoption of cloud computing expands, it is more likely that more and more services will perform mash-ups of data. This development has potential security implications, both in terms of data leaks, and with the sheer number of sources of data that a user may have to pull data from; this, in turn, places requirements on how access will be authorized for reasons of usability [34].

Control issues: Who will be in control of data? As much as cloud computing is expanding, data control issues will become more complex and this will become an important issue internationally. , The December 01, 2008 issue of Star published news indicating that Ottawa had quietly dropped plans to allow the U.S. house a database containing personal information about Canadians who held special drivers licenses for the purpose of crossing the borders. This type of reaction will impose more security and control development on researchers, or the Cloud will never become what its founder's envisioned.

4.8 Data Storage Systems

Data security in cloud computing is a major point of interest for many research papers and projects. The issue is derived from handing customer data to the cloud provider. There are a handful of data distribution models in the cloud environment and almost all of them are adopted from traditional network computing and have been modified to fit into cloud computing. Yet, even after the modifications, there still remain some major security issues unresolved. The point is that cloud providers try to maintain their services' scalability and therefore, keep the file systems as simple and as extendable as possible. Like the two different levels of a scale, this effort decreases the security level, but the problem is how to increase the security level while keeping scalability at the same level or even to improve it.

Network File System (NFS) [37] , is a way to share files in a network in a way that it is perceived to be a local file on the client machine. This makes the file available for all the clients who are able to fake the "real client" identification for the server. As is apparent, the only security is on the authentication server and when this technology is adopted by cloud computing, the situation will become even worse.

Andrew File System (AFS) [38] is a distributed networked file system that uses a set of trusted servers to present a homogeneous, location-transparent file name space to all of the client workstations. This specific file system has some benefits over traditional file systems and NFS. It uses Kerberos for authentication and that makes it sturdy and durable. It also benefits from managing groups and users' dictionary lists and file caching for each client. This also makes file access faster, but because of its characteristics has not yet been adopted by enterprises that possess a vast amount of users such as Facebook and Google. It also maintains a complete copy of each file and some backups on physical servers, which creates security alerts if an attacker is able to successfully bypass Kerberos.

General Parallel File System (GPFS) [39], created and used by IBM, is a shared-disk file system for cluster computers. Since the GPFS uses a centralized management system, it has serious scalability issues and has not adopted by cloud computing technologies enough/to the point that it may be called a usable file system in cloud, despite the fact that IBM is still using the GPFS.

Frangipani [40] is based upon a centralized management system that manages a group of computers and virtualizes a super computer for file storage purposes. This is the same as AFS, as it stores the entire file in ~~a~~ "physical" storage and suffers from the same issue.

Intermezzo [41], the key design is to exploit local file systems as server storage and as a client cache and to make the kernel file system driver a wrapper around the local file system. Intermezzo utilizes the same file treatment as NFS does with the same security issues.

Google File System (GFS)/Hadoop [42], *Distributed File System (HDFS)* [43], are the major distributed file systems in cloud computing. The HDFS is about the most popular adopted file system by major cloud providers. These are the most common algorithms deployed in large scale distributed systems such as Facebook, Google, and Yahoo today.

These file systems use a name node to keep a list of all files in the cloud and their respective metadata (i-node). Along with the fact that the name node has to manage almost all file related operations such as open, copy, move, delete, and update, etc., it may not scale and may potentially cause the name node to become a resource bottleneck.

Another limitation is that the name node is the single point of failure for an HDFS installation [43]. If the name node crashes, the file system will go offline, and that might lead to a huge security issue. Also, if the name node becomes vulnerable to an attacker, data access will become available too.

4.9 Google File Systems / Hadoop File System

According to online information [44], The Google File System (GFS) or Hadoop File System is optimized for Google's core data storage and usage needs (primarily the search engine), which can generate enormous amounts of data that must be retained. The GFS grew out of an earlier Google effort, "BigFiles," developed by Larry Page and Sergey Brim in the early days of Google, while it was still located in Stanford**.**

In GFS, files are divided into chunks of 64 megabytes, which are only extremely rarely overwritten or shrunk; files are usually appended or read. It is also designed and optimized to run on Google's computing clusters, dense nodes that consist of cheap, "commodity" computers, meaning that precautions must be taken against the high failure rate of individual nodes and the subsequent data loss. Other design decisions select high data throughputs, even when it comes at the expense of latency.

The nodes are divided into two types: one Master node and a large number of chunk servers [44]. Chunk servers store the data files, with each individual file broken up into fixed size chunks (hence the name) of about 64 MBs, similar to that of clusters or sectors in regular file systems. Each chunk is assigned a unique 64-bit label, and the logical mappings of files to their constituent chunks are maintained. Each chunk is replicated several times throughout the network, with the minimum being three, but sometimes many more for files that have a high end-in demand or that need more redundancy.

The Master server does not usually store the actual chunks, but rather all of the metadata associated with the chunks, such as the tables mapping the 64-bit labels to the chunk locations and the files they comprise, the locations of the copies of the chunks, what the processes are reading or writing to a particular chunk, or taking a "snapshot" of the chunk pursuant to replicate it (usually at the instigation of the Master server in the event of node failures, when the number of copies of a chunk has fallen beneath the set number). All of this metadata is kept current by the Master server, periodically receiving updates from each chunk server ("Heart-beat messages").

Permissions for modifications are handled by a system of time-limited, expiring "leases", where the Master server grants permission to a process for a finite period of time during which no other process will be granted permission by the Master server to modify the chunk.

The modifying chunk server, which is always the primary chunk holder, then propagates the changes to the chunk servers with the backup copies. The changes are not saved until all chunk servers acknowledge, thus guaranteeing the completion and atomicity of the operation.
Programs access the chunks by first querying the Master server for the locations of the desired chunks; if the chunks are not being operated on (e.g. no outstanding leases exist), the Master replies with the locations, and the program then contacts and receives the data from the chunk server directly (similar to Kazaa and its super nodes).

As opposed to other file systems, GFS is not implemented in the kernel of an operating system, but is instead provided as a user space library.

A GFS cluster consists of a single Master and multiple chunk servers and is accessed by multiple clients [45]. Each of these is typically a commodity Linux machine running a user-level server process. It is easy to run both a chunk server and a client on the same machine, as long as the machine resources permit and the lower reliability caused by running a possibly flaky application code is acceptable. Files are divided into fixed-size chunks. Each chunk is identified by an immutable and globally unique 64 bit chunk handle assigned by the Master at the time of chunk creation. Chunk servers store chunks on local disks as Linux files and read or write chunk data specified by a chunk handle and byte range. For reliability, each chunk is replicated on multiple chunk servers. By default, we store three replicas, though users may designate different replication levels for different regions of the file namespace.

The Master maintains all of the file system metadata. This includes the namespace, the access control information, the mapping from files to chunks, and the current locations of chunks. It also controls system-wide activities such as chunk lease management, garbage collection of orphaned chunks, and chunk migration between chunk servers. The Master periodically communicates with each chunk server in Heart Beat messages to give it instructions and to determine its state.

The GFS client code linked into each application implements the file system API and communicates with the Master and chunk servers to read or write data on behalf of the application. Clients interact with the Master for metadata operations, but all data-bearing communication goes directly to the chunk servers. The POSIX API is not provided and therefore does not need to hook into the Linux Vnode layer. Neither the client nor the chunk server caches file data. Client caches offer little benefit because most applications stream through huge files or working sets that are too large to be cached, and not having them simplifies the client and the overall system by eliminating cache coherence issues [45]. (Clients do cache metadata, however.) Chunk servers do not need to cache file data because chunks are stored as local files and so linuxe's buffer cache already maintains frequently accessed data in memory. According to online information [46], the HDFS provides a framework for storing and processing petabytes of data using commodity hardware and storage.

The Hadoop Distributed File System (HDFS™) is the primary storage system used by Hadoop applications. The HDFS creates multiple replicas of data blocks and distributes them on compute nodes throughout a cluster to enable reliable, extremely rapid computations. The HDFS is a distributed, scalable, and portable file system written in Java for the Hadoop framework. Each node in a Hadoop instance typically has a single data node; a cluster of data nodes form the HDFS cluster.

The situation is typical because each node does not require a data node to be present. Each data node serves up blocks of data over the network using a block protocol specific to the HDFS. The file system uses the TCP/IP layer for communication; clients use RPC to communicate with each other. The HDFS stores large files (an ideal file size is a multiple of 64 MB) across multiple machines. It achieves reliability by replicating the data across multiple hosts, and hence does not require RAID storage on hosts.

With the default replication value of three, data is stored on three nodes: two on the same rack, and one on a different rack. Data nodes may communicate with each other to rebalance data, to move copies around, and to keep the replication of data high. HDFS is not fully POSIX (Portable Operating System Interface for Unix) compliant because the requirements for a POSIX file system differ from the target goals for a Hadoop application. The tradeoff of not having a fully POSIX compliant file system is increased performance for data throughput. The

HDFS was designed to handle very large files, but the HDFS does not provide High Availability [46].

An HDFS file system instance requires one unique server, the *name node*. This is a single point of failure for an HDFS installation. If the name node fails the file system will go offline, and when it comes back online, the name node must then replay all of the outstanding operations. This replay process may potentially take over a half an hour for a big cluster. The file system includes what is called a *secondary name node*, which misleads some people into assuming that when the Primary name node goes offline, the Secondary name node will take over.

In fact, the Secondary name node regularly connects with the Primary name node and builds snapshots of the Primary name node's directory information, which is then saved to local/remote directories. These check pointed images may be used to restart a failed primary name node without having to replay the entire journal of file system actions and the edit log to create an up-to-date directory structure [46].

An advantage of using the HDFS is data awareness between the job tracker and task tracker. The job tracker schedules map/reduce jobs to task trackers with an awareness of the data location. An example of this would be if node A contained data (x, y, z) and node B contained data (a, b, c). The job tracker would schedule node B to perform map/reduce tasks on (a, b, c) and node A would be scheduled to perform map/reduce tasks on (x, y, z). This reduces the amount of traffic that occurs on the network and prevents unnecessary data transfer. When the Hadoop is used with other file systems, this advantage is not always available, and this may have a significant impact on the performance of job completion times, something that has been demonstrated when running data intensive jobs [47].

Another limitation of the HDFS is that it may not be directly mounted by an existing operating system. Getting data into and out of the HDFS file system, an action that often needs to be performed before and after executing a job may be inconvenient. A File system in the User space (FUSE) virtual file system has been developed to address this problem, at least for Linux and some other UNIX systems.

It is probably clear by now that the HDFS is not a general-purpose file system. Instead, it is designed to support streaming access to large files that are written once. For a client seeking to write a file to the HDFS, the process begins with caching the file to the temporary storage locale to the client. When the cached data exceeds the desired HDFS block size, a file creation request is then sent to the name node.

The name node responds to the client with the Data Node identity and the destination block, and the Data Nodes that will host file block replicas are also notified. When the client starts sending its temporary file to the first Data Node, the block contents are relayed immediately to the replica Data Nodes in a pipelined fashion. Clients are also responsible for the creation of checksum files that are also saved in the same HDFS namespace. After the last file block is sent, the Name node commits the file creation to its persistent Meta data storage (in the Edit Log and FsImage files) [48].

File access may be achieved through the native Java API, the Thrift API to generate a client in the language of the users' choosing (C++, Java, Python, PHP, Ruby, Erlang, Perl, Haskell, C#, Cocoa, Smalltalk, and OCaml), the command line interface, and browsed through the HDFS-UI webapp over HTTP [44].

CHAPTER 5: PARTIALLY DISTRIBUTED FILE SYSTEM WITH PARITY

The Partially Distributed File System with Parity (PDFSP) is the protocol that the author has developed as a modification on the existing GFS/HDFS protocol [49]. The PDFSP addresses the three aspects of security (CIA) in data storage and file distribution. A file that will be stored in the cloud will be divided to n chunks (see Figure 5) with specific headers and footers that will be assigned to them by the PDFSP protocol. Each file chunk contains a header with the following information (see Figure 6):

I. 128 bit local deciphering key that is used to construct the complete deciphering key thanks to added bits from other servers.

II. 128 bit remote deciphering key that is used to construct the complete deciphering key for the next chunk in the remote server

III. Address of the next chunk server

IV. Address of the next parity server

V. 128 bit Status Code; used to tag the chunk as original or Parity.

VI. 1024 bits of audit data

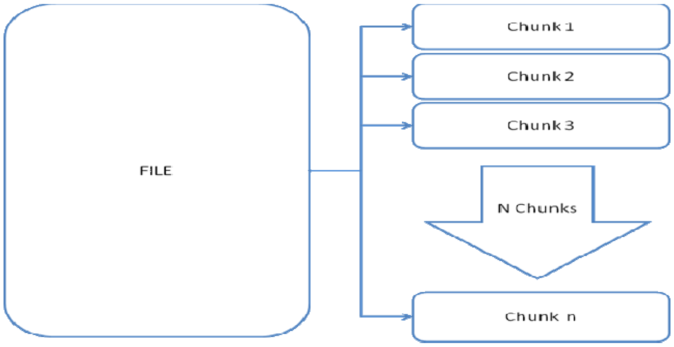

Figure 5. File chunks

Header data is used by the Cloud Management Server (CMS) and the User Management Server (UMS) to find the file chunks, to decipher them, and to add them together (File reconstruction is handled by FRS), and then to rebuild the original file.

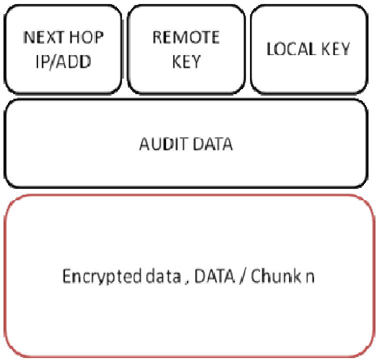

Figure 6. Chunk header

Each chunk is encrypted and the private key to decrypt these chunks will be produced from concatenation of three different keys retrieved from the holding servers. The first part is derived from the previous chunk. If the data chunk is on the server n, the server having the (n-1) chunk will send the remote key to the server n, and the CMS, also sends a part of this key.

The third part is derived from the Customer Client Machine (CCM); These three plus the local key residing on the server n, are concatenated and create the complete deciphering key (see Figure 7).

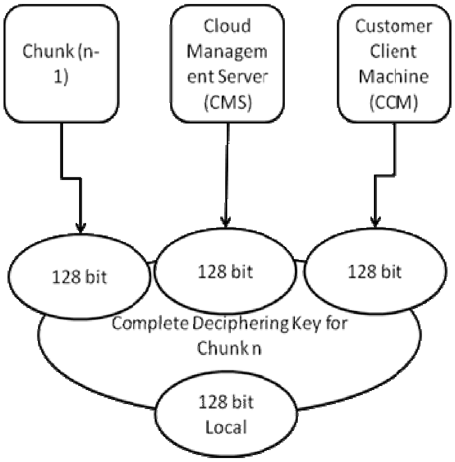
Chunk (n-1)
Cloud Management Server (CMS)
Customer Client Machine (CCM)
128 bit
128 bit
128 bit
Complete Deciphering Key for Chunk n
128 bit Local

Figure 7. Decryption key structure

The integrity and availability of the data is as important as its security. The PDFSP implements parity chunks in order to make sure that the data is available for the user. During the break down process, each chunk may have copies kept in the same manner as that of normal chunks and in the event that the server containing the file is unavailable as a result of an attack or power outage, etc., then the parity system will retrieve the parity chunk instead of the original chunk and will use it to rebuild the original file. Since each file has k copies for parity control, the availability of each chunk in the event that the attacks will be increased; this forces a customized balance between confidentiality and availability characteristics.

5.1 PDFSP Components

The PDFSP Protocol implements the following components:

Client Access Machine (CAM), is a computer that is registered to the cloud service provider as customer's administration machine. This is implemented by hardware dangle that the service provider supplies to the customer who will always be in need of having it connected to the working machine (for administrative purposes). This is the only station that has full control of data. All of the other stations or public users do not have or have only limited access to the data.

User Public Machine (UPM) is a computer that enables normal users to access the Cloud application or data. This machine does not need any specific specifications.

Cloud Management Server (CMS), is the main server that governs the transactions. The CMS is owned and operated by the cloud provider.

File Retrieval Server (FRS), is a server used to construct the file from its chunks using keys from the CAS, the CMS, and local file servers.

5.2 PDFSP Process

The customer uses CAS and sends an "authentication request" to CMS. After checking the dangle and its approval by the CMS, the CAS receives an SID and a list of services available from the server; See Figure 8 below:

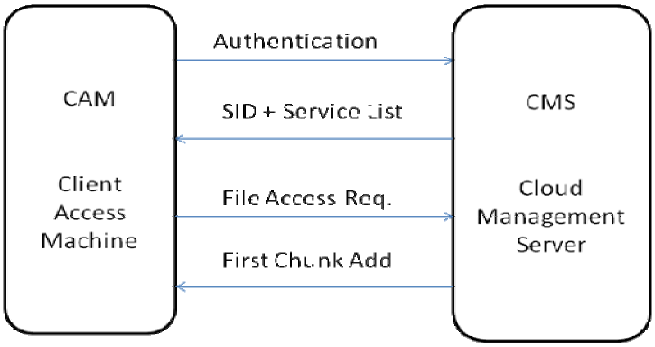

Figure 8. PDFSP Process

Then, the customer sends a file access request to the CMS. Based upon the request, the file access process begins. The goal of the process is to retrieve all of the file chunks and to create an interim copy of file for customer. Finally, after the customer finishes the file manipulation (Read, Update, Delete), the new file chunks are created and stored. The complete file access algorithm is as follows:

I. CAM sends the authentication request to CMS

II. CMS checks and approves CAM Dangle

III. CMS sends SID and Service Lists to CAM

IV. CAM Requests file Access

V. CMS sends the First chunk server address and first chunk (n-1) code part to CAM

VI. CAM sends a File retrieval request, 128 bits of deciphering key, the (n-1) code part, and the first chunk address to FRS (see figure 9)

VII. Based upon an internal algorithm, (Time/Date Dependant – PDFSP RA) CAM's part is translated to the new CCM deciphering code.

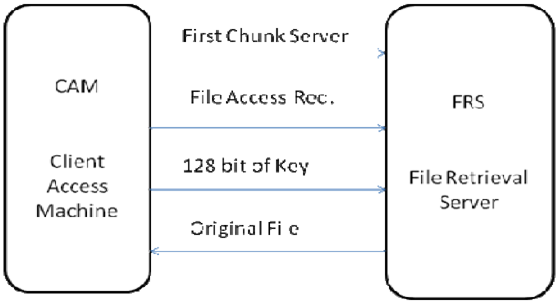

Figure 9. Key transmission

VIII. FRS reads the first chunk from the file server, deciphers the chunk thanks to key parts from CMS, CCM.

IX. FRS refers to the next server by the IP read from the current server.

X. FRS reads the next chunk thanks to key parts from the CCM, the CMS, and the previous chunk, and this repeats until the end of file is reached.

XI. FRS provides the access to the original file for CAM

XII. CAM updates the file and submits it to FRS

XIII. FRS breaks the file down into new chunks and saves the chunks to random servers.

XIV. FRS updates the list in CMS for later access (First chunk address and first deciphering code part)

5.3 PDFSP Security Risk Model

Each file with the weight of W (in Kb) is broken down into n chunks that are distributed over n different servers. Also, each chunk n has a part of its own deciphering key (D_n) and a part of a chunk (n+1) deciphering key (D_{n+1}); it also has the address to the nth chunk (A_{n+1}). A deciphering key is composed of:

$$D_K = D_n + D_{n-1} + D_{CMS +} D_{CCM} \qquad (1)$$

In order to decipher a chunk, the requestor needs to have access to all four parts of the aforementioned equation; The CMS is always running and therefore the chance of finding CMS by a malware or attacker system is**:**

$$P_{CMS} = 1 \qquad (2)$$

The CCM is usually unknown and based upon the CCM connectivity to the cloud. Considering that a malware is trying to gain access to CCM, and having it in mind that a full day (usually malwares are running 24 hours a day) consists of 24x60x60 = 86400 seconds, chance of CCM be available to the malware depends on how many seconds it is connected to the cloud, which is:

$$P_{CCM} = T_{CCM} / 86400 \qquad (3)$$

Where the T_{CCM} is the total connected time for the Customer Client Machine in the scale of seconds.

Given that the total number of the servers implemented by the cloud provider is N_s, it is necessary to select the correct pairs having both D_n and D_{n-1}, since the two chunks are distinct when they are used; therefore, there is no need to re-use them. Also, Let b is the number of the servers with no file. The total number of combinations of n chunks taken two at a time is:

$$C_n = \{(N_s - b) 2-2\} / \{ N_s! / (N_s -2)!2!\} \qquad (4)$$

Where denominator is the number of all choices. The numerator is for picking the first server to choose carefully from the ones with a chunk in them. The chance of gaining access to all the necessary items for deciphering a chunk will be:

$$P_{FA} = \{(N_s - b) 2-2\} / \{N_s! / (N_s -2)!2!\} \text{ X } \{T_{CCM} / 86400\} \qquad (5)$$

In the event of deciphering a chunk, the next chunk server and D_{n+1} will be accessible thanks to the retrieved data, and therefore:

$$P_{n+1} = 1 \qquad (6)$$

Where $$b = N_s - n - K \qquad (7)$$

$$n + K < N_S \text{ and } Ns > 2 \qquad (8)$$

Substituting “b” with its equal value above will lead to the final formula of:

$$P_{FA} = \{T_{CCM}\,(n+K-1)\} \,/\, \{21600 \text{ x } N_s\,(N_s - 1)\} \qquad (9)$$

Where**:**

- $P_{FA:}$ Probability of retrieving a file by an attacker
- n: number of chunks per file
- T_{CCM}: Total time in second that CCM is connected to the cloud
- N_s: Total number of servers used for file storage
- K: Quantity of parity controls

Availability of a file for user, in case of server failure or malware attack, depends on quantity of parity chunks. When there are K<n parity chunks, it means there are file chunks with no backup at all and this is a very risky, but budget friendly situation. In case of having n= K there will be a full copy of whole file (all chunks have a backup copy). Total number of “full backup copies” of a specific file could be determined by (K/n). The fraction of total number of backups to the whole file chunks will be (K/ (n+K)). Having in mind that availability of a server in provider’s system is defined by $P_{p,}$ chance of successful file retrieval after any malware attack/ server failure will be:

$$A_I = K\,P_p \,/\, (n+K) \qquad (10)$$

Where:

- K: Quantity of parity control chunks

- $P_{p:}$ Availability of provider's File servers
- n: number of chunks per file

5.4 PDFSP Security Risk Model Analysis

The flexibility and simplicity of the presented model empowers the cloud provider to balance budget with security needs. Figure 10 assumes that the Customer Client Machine is connected to the cloud for normally six hours of working time and the file chunks are 10 per file with two parity chunks. Vertical axis shows Risk factor while horizontal axis counts the number of server available for cloud provider. As depicted, when the quantity of servers rises, the security increases as well, but the rate of this increment decreases at the same time.

When increasing the number of servers from 15 to 20 (five servers) raises the security about 23%, with the same step of increment of between 48 and 53 only increases the factor by 0.08%. In both cases, the cloud provider should expend the same amount of budget on adding five more servers, but the results will be extremely different.

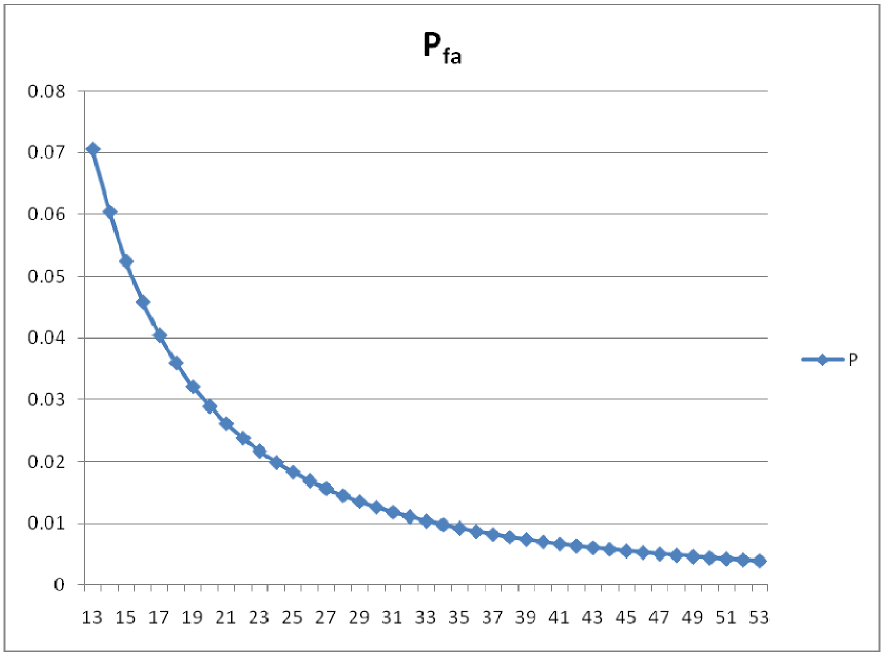

Figure 10. Risk factor vs. N_s

According to formula (9), it is derived that the longer the time that the CCM is online, the higher the number of chunks and parity chunks per file, leading to a higher risk factor. Figure 11 displays the relationship between (K+ n) and the corresponding risk factor.

In studying the file retrieval chance formula (10), it reveals that in the event that there is no parity chunk already set, after a failure, it is impossible to retrieve the file. Also, the more the parity chunks are produced, the higher the chance of file retrieval will be.

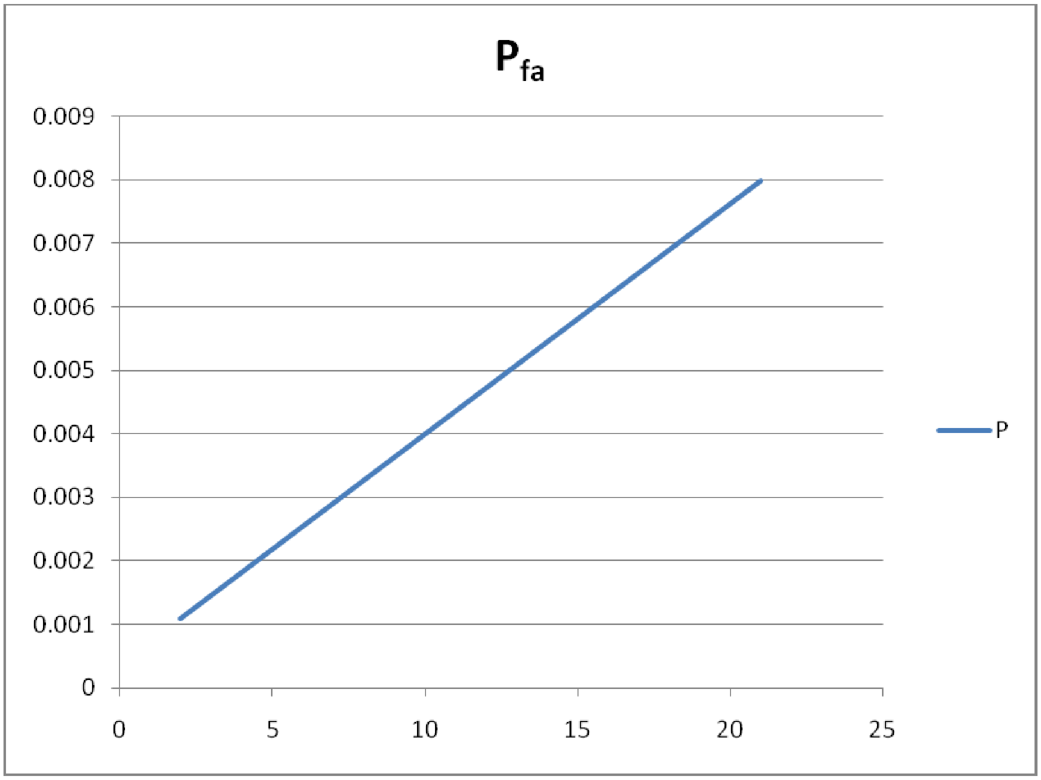

Figure 11. Risk factor vs. (n + K)

Another aspect of the model considers the number of chunks per each file. This is where the server availability factor of the cloud provider plays a very big role. Figure 12 depicts the effect of increasing the number of parity chunks in increasing the chance of file retrieval after a failure or malware attack to a file server. It is obvious that security demands of different customers vary, and this emphasizes the need of an analytical study for each customer and the need to arrive at the balanced point between n and K based upon the available N_s and budget.

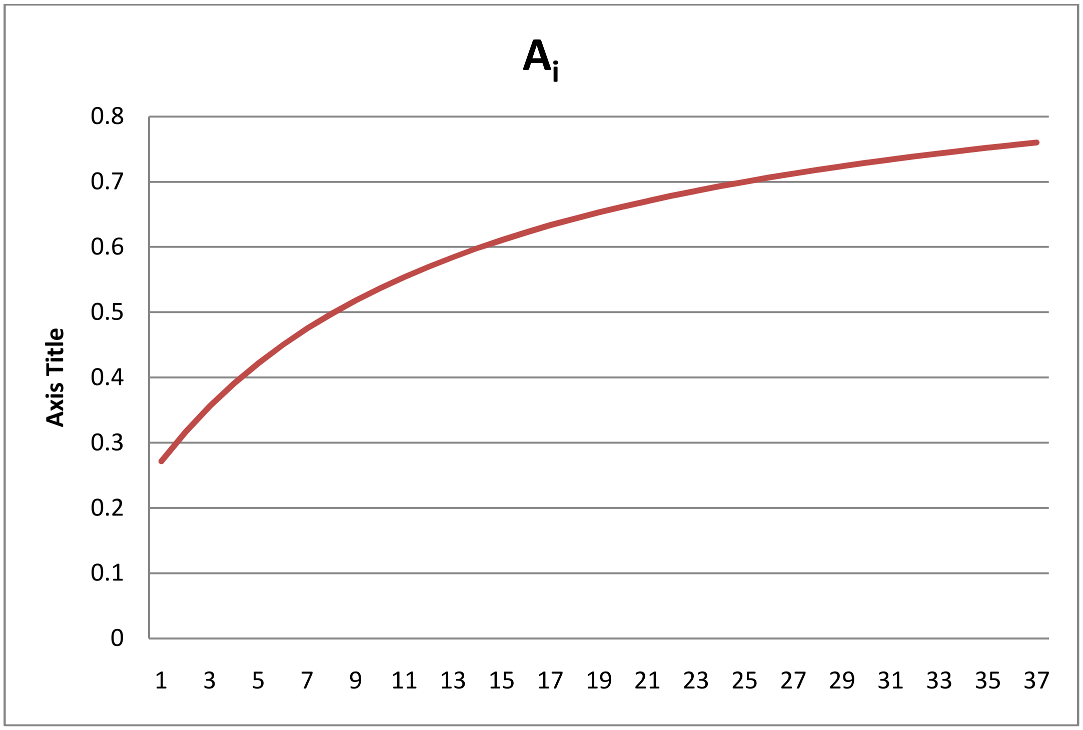

Figure 12. File retrieval chance vs. K

Using the PDFSP assigns the following values to the cloud computing environment:

Every time a file is accessed by the customer, new chunks with new sizes are produced; these chunks are completely different from previous chunks and are stored in new randomly selected servers. This lowers the security threat by eliminating possible re-access to the file by the provider itself or anyone who has already hacked into the protocol.

Three distinctive, physically separated servers are involved in the file retrieval process; this also lowers the chance of success for any attacker. The PDFSP implements a time dependant algorithm that means that at any specific time, the algorithm that is fed by the time string value creates a new decoding password, and again lowers the attackers chance to gain access to any file.

The parity system assists in retrieving data in the event of power outages, system failures, and any other causes that may remove the chunk server from the service line that maintains the specific level of availability for files.

The flexibility of the model helps cloud providers manage their budget by keeping a balance between confidentiality, availability, and integrity according to each specific customers' needs.

The PDFSP is running on the currently available TCP and therefore is not forcing a huge change in the storage process, again saving the budget.

Securing the cloud data increases the prospective customers' confidence in the cloud and therefore more revenue will be generated for cloud providers.

CHAPTER 6: CONCLUSION

Cloud computing is developing rapidly and is commonly believed to be the future of the computation world. In this way, there are major concerns such as security that need to be addressed thoroughly and in depth in order to boost this development. An analysis of the current situation shows that the security level of present solutions is not at the level that would attract new enterprises and convince the ones already studying the technology to migrate from traditional computation technology to cloud computing.

Different file systems have addressed the security issue by now, but they do not seem to be a convincing solution to the problem. Even the major systems, such as GFS and HDFS, both already in use by the largest providers such as Google, appear to be incomplete. In GFS/HDFS architecture, issues like that the Master server stores all the metadata associated with the chunks and similar to this, lead the whole system to be vulnerable to attacks and failures. Attackers just need to access to Master server/node server to gain access to data An HDFS file system instance requires one unique server, the name node. This is a single point of failure for an HDFS installation. If the name node goes down, the file system is offline and this reduces the system's availability rate.

The author's proposed model, the Partially Distributed File System with Parity Chunks, addresses all three aspects of security, including Confidentiality, Integrity, and Accessibility (CIA). The model is designed in a way that is flexible and customizable to fit into any condition and any specific customer need while keeping the budget at an optimum level. Saving the budget at the same time it supports green technology; with optimum number of file servers.

7. References

[1] Knorr E., Grumman G., (2008), “What Cloud Computing Really Means”, Info World, [Accessed 01-06-2011]: http://www.smartchinateam.com/downloads/cloud_computing_intro.pdf

[2] David C. W., (2010), “Burgeoning Clouds: Cloud Computing Will Mean Outsourcing Government Information Technology to a New Level”, Southeastern Louisiana University Technical Journal.

[3] Ramakrishan R., (2010). “What does it Mean to Enterprises?”, [Accessed 08-09-2010]: http://www.cumulux.com/Cloud%20Computing%20Primer.pdf

[4] Clarck D. , (2008), “Enterprise’s Security: The Managers Defense Guide”, Adison-wesley Information Technology Series.

[5] Mcmannus R., (2008), “More Amazon S3 Downtime: How Much is Too Much?”, [Accessed 02-24-2011]: http://www.readwriteweb.com/archives/more_amazon_s3_downtime.php

[6] Keith A. Watson, (2008), “Security Management Practices”, [Accessed 02-27-2011]: http://www.purdue.edu/securepurdue/docs/training/Security+Management+Practices.pdf

[7] Halton G., Deepak S., (2009), “Cloud Computing Essay”, [Accessed 12-04-2010]: http://www.scribd.com/doc/23743963/Cloud-Computing-Essay

[8] Hoffman K., (2009), “Azure Changes Difference Between Web Hosting and Cloud Computing”, [Accessed 09-18-2010]: http://linux.sys-con.com/node/1095058

[9] Gil P., (2010), “What is Cloud Computing”, [Accessed 01-22-2011]: http://netforbeginners.about.com/od/c/f/cloudcomputing.htm

[10] Web site Admin, (2006). “Software as a Service (SaaS) “, [Accessed 09-25-2010]: http://searchcloudcomputing.techtarget.com/definition/Software-as-a-Service

[11] Souter B., (2009), “Who Is the User for Cloud Computing?”,[Accessed 12-11-2010]: http://www.sutor.com/newsite/blog-open/?p=4548/21

[12] OrBytes Website Admin, (2009). “Cloud Computing”, [Accessed 11-08-2010]: http://www.orbytesolutions.com/services/index.php?option=com_content&view=article&id=55&Itemid=40

[13] Waxer B.,(2009), “The Benefits of Cloud Computing”, [Accessed 12-21-2010]: http://www.webhostingunleashed.com/features/cloud-computing-benefits/

[14] Dans E. , (2011), “Benefits and Disadvantages of Cloud Computing”, [Accessed 03-14-2011]: http://algramrandomramblings.blogspot.com/2011/01/benefits-and-disadvantages-of-cloud.html

[15] Kynetix Technology group, (2009), “Cloud Computing Strategy Guide”, [Accessed 12-02-2010]: https://sites.google.com/site/cloudmanual/success-factors

[16] Yachin D., (2009), “It is Time for Stormy Weather”, IDC Emerging Technologies, [Accessed 09-21-2010]: http://www.slidefinder.net/I/IDC_Cloud_Computing_IGT_final/16324826

[17] Anderson J., Rainie L., (2010), “The Future of Cloud Computing”, [Accessed 04-29-2011]: http://www.pewinternet.org/Reports/2010/The-future-of-cloud-computing.aspx

[18] McKendrick J., (2011), “Loud Divide: Senior Executives Want Cloud, Security and IT Managers are Nervous”, [Accessed 04-20-2011]: http://www.zdnet.com/blog/service-oriented/cloud-divide-senior-executives-want-cloud-security-and-it-managers-are-nervous/6484

[19] IDC data survey source, (2008), IDC Enterprise Panel, survey number=244

[20] Lynch D., (2008), “New Security Issues Raised by Server Virtualization”, [Accessed 03-01-2011]: http://www.itworld.com/virtualization/59445/new-security-issues-raised-server-virtualization

[21] Petri D., (2009), “What You Need to Know about Securing Your Virtual Network”, [Accessed 09-14-2010]: http://www.itworld.com/virtualization/59445/new-security-issues-raised-server-virtualization

[22] Suuburah J. , (2010), “Security Issues in Cloud Computing”, [Accessed 02-07-2011]: http://www.computerweekly.com/Articles/2010/01/12/235782/Top-five-cloud-computing-security-issues.htm

[23] Gohring N., (2008). “Amazon Web Services (AWS)”, [Accessed 08-21-2010]: http://aws.amazon.com

[24] Raphael J., (2009), “Google Docs Glitch Exposes Private Files”, [Accessed 06-08-2011]: http://www.pcworld.com/article/160927/google_docs_glitch_exposes_private_files.html.

[25] McMillan R., (2007), “Salesforce.com Warns Customers of Phishing Scam”, [Accessed 01-10-2011]: http://www.pcworld.com/businesscenter/article/139353/salesforcecom_warns_customers_of_phishing_scam.html.

[26] Miller B., (2005), “Security Evaluation of Grid Environments”, [Accessed 02-04-2011]: https://hpcrd.lbl.gov/HEPCybersecurity/HEP-Sec-Miller-Mar2005.ppt.

[27] Bigsey A., (2009), “Cloud Computing and Impact on Digital Forensic Investigation”, [Accessed 05-19-2011]: http://www.zdnet.co.uk/blogs/cloud-computing-and-the-impact-on-digital-forensic-investigations-10012285/cloud-computing-and-the-impact-on-digital-forensic-investigations-10012286/

[28] Prez J., (2008), “Extended Gmail Outage Hits Apps Admins”, [Accessed 06-27-2011]: http://www.computerworld.com/action/article.do?command=viewArticleBasic&articleId=9117322.

[29] Amazon S3 Team, (2008), “Amazon S3 Availability Event”, [Accessed 12-15-2010]: http://status.aws.amazon.com/s3-20080720.html.

[30] Tubanos A., (2008), “FlexiScale Suffers 18-Hour Outage”, [Accessed 11-21-2010]: http://www.thewhir.com/web-hosting-news/103108_FlexiScale_Suffers_18_Hour_Outage.

[31] Masuoka R., Molina J., Chow R., Jacobson M., (2009), “Outsourcing Computation without Outsourcing control”, [Accessed 03-07-2011]: http://markus-jakobsson.com/papers/jakobsson-ccsw09.pdf

[32] www.Animoto.com

[33] Time Machine, “The New York Times”, [Accessed 01-10-2011]: http:/timesmachine.nytimes.com

[34] ISO, (2008), “ISO 27000 Directory”,[Accessed 01-02-2011]: http://www.27000.org/iso-27005.htm

[35] Gellman R., (2009), “Privacy in the Clouds: Risks to Privacy and Confidentiality from Cloud Computing”, [Accessed 03-11-2011]: http://www.worldprivacyforum.org/pdf/WPF_Cloud _Privacy_Report.pdf

[36] Kerbs B., (2008), “Lithuania Weathers Cyber Attack, Braces for Round 2”, [Accessed 02-10-2011]: http://voices.washingtonpost.com/securityfix/2008/07/lithuania_weathers_cyber_attac_1.html

[37] Shepler S., Callaghan B., Robinson D., Eisler, M., and Noveck, D., (2003), ” Network File System (NFS) Version 4 Protocol. Request for Comments 3530”, [Accessed 03-21-2011]: http://www.ietf.org/rfc/rfc3530.txt

[38] Howard J., Kazar M., Menees S., Nichols D., Satyanarayanan M., Sidebotham R., and West M., (1998), “Scale and Performance in a Distributed File System ”, [Accessed 12-06-2010]: http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.71.5072

[39] Schmuck F., and Haskin R., (2002), “A shared-disk file system for large computing clusters. Proceedings of the 1st USENIX Conference on File and Storage Technologies “USENIX Association, p. 19

[40] Thekkath C., Mann T., and Lee E. Frangipani, (1997). “A Scalable Distributed File System”, ACM SIGOPS Operating Systems Review 31, 5, 224–237

[41] Braam P., Callahan M., and Schwan P., (2000), “The Intermezzo File System”, Proceedings of the 3rd of the Perl Conference, O¨ı£¡Reilly Open Source Convention, Citeseer

[42] Ghemawat S., Gobioff H., and Leung S., (2003), “The Google FileSsystem”, *SIGOPS Oper. Syst. Rev. 37*, 5

[43] Borthakur D, (2007), “The Hadoop Distributed File System: Architecture and Design “, the Apache Software Foundation, [Accessed 01-18-2011]: http://hadoop.apache.org/common/docs/r0.18.0/hdfs_design.pdf

[44] http://en.wikipedia.org/wiki/Google_File_System

[45] Ghemawat S., Gobioff H., and Leung S., (2003), “The Google File System” [Accessed 03-20-2011]: http://labs.google.com/papers/gfs.html

[46] http://en.wikipedia.org/wiki/Apache_Hadoop

[47] Xie J., Yin S., Ruan X., Ding Z., Tian Y., Majors J., Manzanares A., and Qi X. , (2010), “Improving MapReduce Performance through Data Placement in Heterogeneous Hadoop Clusters”, [Accessed 05-17-2011]: http://www.eng.auburn.edu/~xqin/pubs/hcw10.pdf

[48] http://www.ibm.com/developerworks/linux/library/l-hadoop/

[49]www.Mogulus.com

[50]www.Elililly.com

[51] Cloud Security Alliance, (2009), “Security Guidance for Critical Areas of Focus in Cloud Computing”, [Accessed 02-27-2011]: http://www.cloudsecurityalliance.org/guidance/csaguide.pdf.

[52] Yesn Y., (2008), “Why Google Apps is not Being Adopted”, [Accessed 03-09-2011]: http://money.cnn.com/2008/08/19/technology/google_apps.fortune/index.htm.

Appendix A

Some Major Cloud Computing Providers

- Telstra
- Vodafone
- Akamai
- Limelight
- BT
- AT&T
- Go Grid
- Amazon Web Services
- Mosso
- Google
- Cloud Works
- Windows Azure
- Verizon Wireless
- Ping Identity
- Tricipher
- Data Direct
- Strike Iron
- Cisco WebEx
- Live Meeting
- Microsoft Office Live
- Sales Force

- Workday
- Blogger
- Right Now
- Flicker
- Force.com

Appendix B

Cloud user Examples

Mogulus: Mogulus is a live broadcast platform on the Internet. As a cloud customer, producers may use the Mogulus browser-based studio application to create LIVE, scheduled, and on-demand Internet television for broadcast anywhere on the web through a single player widget. Mogulus is entirely hosted on the cloud. A report from their website [49], says that on election night they amped to 87,000 videos at 500 Kbps, an almost 43.5 Gbps data rate!

Animoto: Animoto is a video rendering and production house with services available on the Internet. With their patent-pending technology and high-end motion design, each video is a fully customized orchestration of user-selected images and music in several formats, including DVD. Animoto is entirely hosted on the cloud. They also have released a Facebook app for users to easily render their photos into MTV-like videos. The first time this app was released, they ramped from 25,000 users to 250,000 users in just three days; the rate of users has been 20,000 signing ups per hour, forcing them to go from 50 to 3,500 servers in merely five days (two weeks later they scaled back to 100 servers) [32].

Eli Lilly: Eli Lilly is the 10th largest pharmaceutical company in the world. They moved their entire R&D environment to the cloud. The results for Eli Lilly were reduced costs, Global access to R&D applications, and Rapid transition as a result of VM Hosting. The vital result is that the time to deliver new services has been greatly reduced. This time has been scaled down for new servers from 7.5 weeks to just three minutes! Additionally, for new collaborations the time was scaled down from eight weeks to five minutes! [51].

New York Times: The Time Machine is a news archive of the New York Times available in pdf form on the Internet for newspaper subscribers; it is entirely hosted on the cloud. The Time Machine needed infrastructure to host several terabits of data, and this was rejected by internal IT as a result of cost and difficulty, business owners had the data up on the cloud for $50 over one weekend! [33].