International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2014): 5.611

A Review on Power Saving & Security Mechanisms in Cloud Computing

Varun Pankaj Joshi[1], Vina M. Lomte[2]

[1]Department of Computer Engineering, R.M.D. Sinhgad School of Engineering

[2]Assistant Professor & Head, Department of Computer Engineering, R.M.D. Sinhgad School of Engineering

Abstract: ***Cloud Computing is a distributed technology that influences sophisticated technology innovations which are highly scalable & resilient and can be used in a variety of powerful ways. The primary focus of the Cloud is the delivery of remote computing resources over the Internet in an affordable and reliable manner. The on-demand nature of cloud services result in the need of energy efficient servers which can satisfy the performance expectations of a cloud user. Servers catering to a constantly increasing user base find it difficult to conserve power due to the random and on-demand nature of the cloud. While cloud computing puts forth an economic ease for providers as well as users, it poses striking security challenges to preserve user data and maintain their trust over this technology. In spite of significant research and development, loopholes still exist in the polices of Cloud Computing. The fundamental aim of this paper is to explore various power saving and security mechanisms for Cloud Computing. Several power saving polices are studied and their efficiency is compared. Different security mechanisms were taken into consideration and discussed. Increasing need for on demand computing services on various platforms brings a new level of attention to energy efficiency and security in the world of the cloud.***

Keywords: energy-efficient control, cloud security, cloud computing, power saving, server management

1. Introduction

The past couple of decades have seen the powerful business concept of outsourcing services and utility computing evolve in parallel. The IT industry soon realized the profitable impact of this parallelism and branded the term „cloud computing" which is now more than just a technology trend. Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction [1]. The cloud architecture comprises of three models: Infrastructure as a Service (IaaS), Platform as a Service (PaaS) & Software as a Service (SaaS). Amazon Web Services, a popular example of an IaaS is a collection of services such as the Amazon Elastic Compute Cloud (EC2), Amazon Simple Storage Service (S3) etc. which allow users to use resources for their computing and storage needs. [2]

The on-demand nature of the cloud leads to uncertain workloads and the major requirement of being available to users anywhere, anytime for which more resources need to be allocated. This leads to over-provisioning and redundancy which are common in a traditional Operating System [3]. Most data centers tend to suffer from idle-times and under-utilization since demand cannot be predicted and might be unavailable for days, or even weeks. As of 2010, 82% [4] of the servers in major data centers went un-utilized. On average, a server will remain busy for only 20-30% of the time.

Cloud Computing is predicted to grow substantially over the next few years, with major corporations moving themselves over to the cloud. Forrester has predicted the cloud industry to reach over $240 billion in 2020 [5]. This also results in substantial power consumption in turn leading to huge amount of carbon emissions [6], [7]. Shutting down the servers when idle might seem like a direct and ideal way to save power, which is not the case since it is a violation of the Quality of Service(QoS) for Cloud Computing [8].

To avoid this scenario, an „N Policy", brought forward by Yadin and Naor [9] is extensively adopted in fields such as communication systems etc. A server will turn on only when jobs in the queue are greater than or equal to a pre-defined value of „N". However, this method may result in degraded performance when „N" is a large value leading to the server staying in power-saving mode. Three power saving policies defined by Chiang et al. [10] focus on switching the server between idle, sleep and power-on modes are discussed in the following sections.

Some individuals believe that cloud is storing your data on someone else"s computer [11]. However, numerous security mechanisms are at play to ensure that our data remains our data. Authentication is another important aspect of the cloud. While there are a lot of security mechanisms available to address the security issues, none of them is fool proof. Since data can be transferred over a wireless medium it is vulnerable to unauthorized access and modifications. It is important to establish the best in class security policy to make sure users trust the cloud with their private information.

The rest of this paper is structured as follows. Section 2 gives an overview about the related work in power saving and security methods, Section 3 discusses power-saving policies suggested by Chiang [10] and Project Natick, Section 4 compares these polices, Section 5addresses cloud security mechanisms such as Public Key Infrastructure (PKI), Single Sign-On (SSO) etc. Finally, the conclusions are presented in Section 6.

2. Related Work

Cloud Computing is being studied extensively over the recent past to make it as efficient, reliable and affordable as

possible. Virtual Machines are optimized to provide the best performance while minimizing power consumption. Various algorithms are under development to analyze and optimize the server performance. A Microsoft Research project has recently deployed a Cloud Data Center in the sea.

A physical server in a data center is busy around 20-30% of the time on average. Hence a number of physical servers working at this rate lowers the efficiency of the entire data center. Power consumption can be significantly reduced by migrating virtual machines to one server in order to accomplish given job. This reduces the need of two physical servers working to complete the same tasks. Server consolidation and proper virtual machine migration also reduces the overall cost of hardware. [12]

In [13], A virtual machine allocation algorithm was presented by Huang, Li and Qian which aims at minimizing the overall power consumption. In the event of a job arrival, the algorithm checked the resource usage status for each server, and chose the most suitable server to allocate the task, thereby reducing actual number of physical servers working at any given instant.

The Microsoft Research Project deployed an underwater data center with the aim of utilizing natural cooling offered by water. This drastically reduces the need to provide external cooling and thus reduces the carbon footprint of the facility.The experimental prototype, „Leona Philpot" [14] was operated underwater approximately one kilometer off the Pacific Coast from August to November 2015. The 38,000-pound container had a computing power equivalent to 300 personal computers. Such deployment brings the datacenter closer to the masses and hence increases the speed of data transfer. The time taken to build the underwater data center takes around 90 days, much less than datacenters which consume real estate and take about 24 months to become operational.

Data centers are the prime locations which require high amount of cooling since they house a large number of physical machines. Processers dissipate heat and too much of heat may cause damage to the servers.

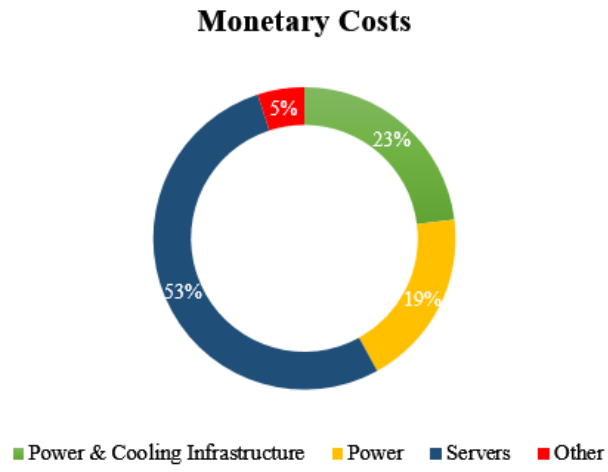

Figure 1: Cooling and power costs constitute over 40% of the total cost [15]

Google has about 16 data centers worldwide. As of 2013, Google data centers house around 900,000 servers which use about 260 million watts of power which accounts to 0.01% of the global energy. Such large amounts of power can power up to 200,000 homes.

The power consumption of a large data center in the United States can power an entire small town. Data centers account to about 17% of the carbon footprint. Data centers which are older than 7 years do not comply with Green Computing norms. The average life span [16] of a data center is considered to be 9 years.

3. Power Saving in the Cloud

3.1. The ISN policy

A typical cloud consists of physical servers, a job dispatcher or scheduler and virtual machines (VMs) running on the servers. The job dispatcher in the policies discussed by Chiang [9] is used to identify a job request and forward it to a queue maintained by the corresponding VM manager that can satisfy its Quality of Service (QoS) standards and meet its specific requirements. If no job requests are available, the server goes into idle state. For a cloud system, a server switches between busy and idle state due to random job arrivals. A server cannot switch from busy to sleep directly, it has to remain idle for a certain about of time [17].

Step1. Incoming jobs are serviced while a server is in a busy mode. A server can end its busy mode when the current job requests have been finished and the queue becomes empty.
Step2. A server stays in an idle mode and waits for job arrivals before switching into sleep mode.
Step3. If a job arrives during an idle period, a server can switch into a busy mode and starts to work immediately. Then, a server begins an idle period until all job requests have been successfully completed.
Step4. If there has no job arrival, a server switches into a sleep mode when an idle period expires.
Step5. A server remains in a sleep mode if the number of jobs in the queue is less than the pre-defined N value. Otherwise, a server switches into a busy mode and begins to work. The two cases of starting a busy mode are:
Case 1: starting a busy mode when a job arrives in an idle mode;
Case 2: starting a busy mode if the number of jobs in the queue is more than the N value.

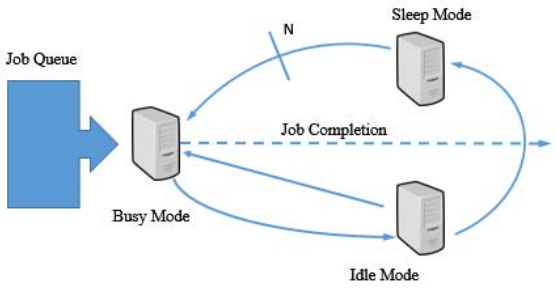

Figure 2: A system with ISN policy

3.2. The SN & SI policies

A server remaining in the idle state still consumes 60% power at its peak power [4]. To reduce power consumption, Chiang et al. propose that a non-idle mode be used where the

server can switch between Busy and Sleep modes only. Here, instead of going to idle state when queue is empty, the server switches into a sleep state.The switching restriction is given by the pre-defined N value.

Step1. A server switches into a sleep mode immediately when no job is in the system.
Step2. A server stays in a sleep mode if the number of jobs in the queue is less than the N value; otherwise, a server switches into a busy mode and begins to work.

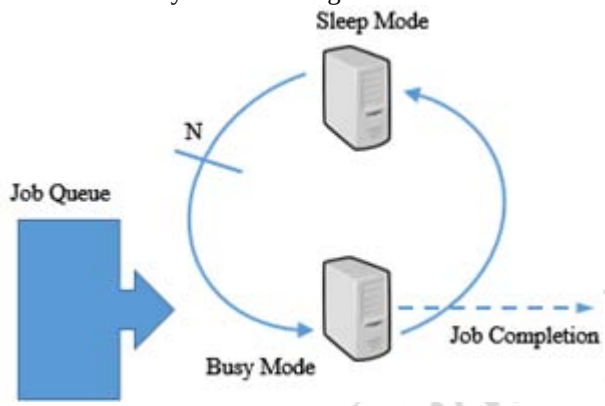

Figure 3: A System with the SN policy

Alternatively, the SI policy proposed in [10] switches the server from Busy to Idle depending upon whether a job is available or not.

Step 1: A server immediately switches into a sleep mode instead of an idle mode when there has no job in the system.
Step 2: A server can stay in a sleep mode for a given time in an operation period. If there has no job in the queue when a sleeping time expires, a server will enter into an idle mode. Otherwise, it switches into a busy mode without any restriction and begins to work.

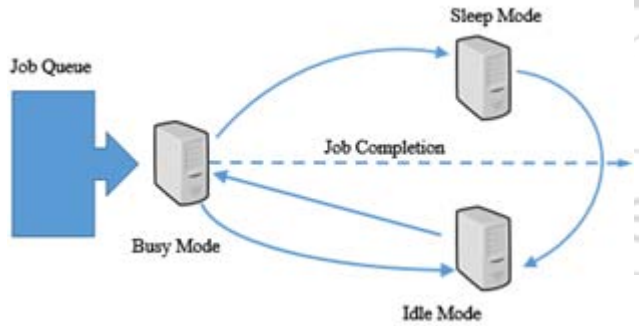

Figure 4: A System with the SI policy

3.3. Relevant Mathematics

A larger controlled N value gains more power saving but results in excessive delay. A smaller N value reduces delay times but leads to short operational cycle. The power consumption overhead cannot be ignored. Operational costs, system congestion costs and performance degradation costs were considered in the cost function [10]. Corresponding cost functions are as follows:

Table 1: Cost Notations

Notation	Description
C_0	Power cost when a server is in busy mode per unit time
C_1	Power cost when a server is in idle mode per unit time
C_2	Power cost per service rate per unit time
C_3	Power cost when server is in sleep mode per unit time
C_4	Startup cost incurred by activating a server
C_5	Cost incurred by jobs waiting in queue
C_6	Congestion management cost per unit time

C_5mainly indicates performance penalty cost used to compensate for user delay. Congestion management cost is spent to manage jobs according to a scheduling algorithm (eg: FCFS, LCFS). The mean length of operational period is denoted by E[C] (to be found at [17],[18],[19]) and it helps estimate startup cost. A response time guarantee is one the most important performance parameter in a green cloud environment since no user would want to experience unusual delay caused by power saving. Hence, both the waiting time and execution time are considered. Stating the problem mathematically,

Minimize F_c
Where Fc = Fc (μ, N)
$= C_0P_B + C_1P_I + C_2\mu + C_3P_S + C_4/E[C] + C_5W + C_6L$(1)
$0 <= \rho <= 1$ (i)
W <= x (ii)
P_B, P_I, $P_S$denote the probabilities that a server is in Busy, Idle or Sleep mode respectively [10]. „x" is the response time guarantee and W is wait time.

3.4. Efficient Green Control (EGC) Algorithm

Input:
1. An arrival rate λ.
2. Upper bound of the server rate and the waiting buffer, denoted by μu and Nb
3. Cost parameters [Co, C1, …, C6]
4. A response time guarantee x.
5. System parameters{Θd, Θs} used by the ISN policy
6. System parameter {k} used by the SN policy
7. System parameters {Θ, N=1} used by the SI policy

Output: μ*, N* and Fc (μ*, N*)

Step1. For i = 1; i = u; i++;
Set μi ← a current service rate;
Step 2. For j = 1; j = b; j++ ;
Set Nj ← a current N parameter;
Step 3. Calculate the system utilization.
If the current test parameters satisfy the constraint of (i) $0 \leq \rho \leq 1$, then
Calculate the response time;
Else
Return to step 1 and begin to test a next index;
End
Step 4. If the current test parameters satisfy the constraint of (ii) $W \leq x$, then
Record the current joint values of (μi, Nj) and identify it as the approved joint parameters;
Else
Return to step 1 and begin to test a next index;
End

This algorithm is further applied to ISN, SI & SN policies for experiments by Chiang et al. [10]

3.5. Project Natick

Project Natick is a research project by Microsoft to estimate the feasibility of underwater datacenters [20]. The project seeks to understand the benefits and obstacles associated with underwater datacenters.*Leona Philpot*, the experimental undersea datacenter was operated on the seafloor approximately one kilometer off the Pacific coast of the United States from August 2015 to November 2015. The project reflects Microsoft‟s ongoing efforts for cloud solutions that offer quick deployment, rapid provisioning, lower costs, high responsiveness.

Such datacenters have the ability to be deployed rapidly, from start to finish in 90 days. This enables cloud providers to respond quickly to market demand and help restore connectivity in times of natural disasters or crisis. About 44% [21] of the world‟s population lives within 150 kilometers of the coast. A subsea datacenter can help reduce proximity of this huge amount of population to datacenters and help decrease data latency. As of now, this datacenter was powered by land-based power grid.

The datacenter is designed to last up to 5 years which is the intended lifespan of the computers in the datacenter. After 5 years, the datacenter is to be retrieved and the computers are to be replaced before deployment into the sea. A Natick datacenter is designed to last 20 years before it needs to be recycled.

The objective is to create a sustainable datacenter which uses green energy, providing customers with additional options to meet their own requirements. Natick datacenters are fully recyclable. If coupled with renewable resources, they could be truly zero emission (No waste from power generation or human intervention). As the end of Moore‟s Law is near, rate at which servers are upgraded with new hardware is likely to slow down. They also do not consume water for cooling purposes.

4. Comparison of Algorithms

There are two main benefits of adopting the ISN policy. It gives arrival jobs more possibilities of getting services without latency. Secondly, the total startup cost can be reduced since a server remains active at regular service rate. The ISN policy considered here follows deterministic (fixed) time.

In the SN policy the server directly switches to sleep mode instead of entering an idle mode. To avoid switching too often, the sleep-busy states are controlled by a predefined N value. This significantly reduces the amount of power wasted in keeping the server idle and also the N value controls switching. However, if number of jobs remain less than N for a long time, the server may remain asleep and the jobs would not be serviced at all.

On the other hand, if no mode-switching restriction is imposed on the server, the server will slip into sleep mode right away rather than being idle when no jobs are in queue. When sleeping time expires, the server switches to idle state if no job is available else it starts working immediately. In such a scenario a job may experience high amount of latency if the sleep period is set to high value. Also if no jobs arrive for significant amount of time, the server may keep waking up from sleep and remain idle, thus wasting power.

Table 2: Comparing ISN, SN, SI algorithms

	ISN	SN	SI
A server goes into a sleep mode immediately when a queue becomes empty		Y	Y
Switching into a busy mode depends on the number of jobs in a queue	Y	Y	
An idle server is not allowed in a system		Y	
Applying the mode-switching control	Y	Y	
Having exponential service times	Y		
Having general service times			Y
Having deterministic (constant) idle times	Y		

The proposed power-saving policies can effectively reduce cost, especially when the arrival rate is low. The performance and cost improvement is shown in the figure 5. According to results, applying the SI policy can obtain better response times when arrival rate is low. A cloud provider whose main objective is to reduce cost, implementing the SN policy is a better option while dealing with variable arrival rates.

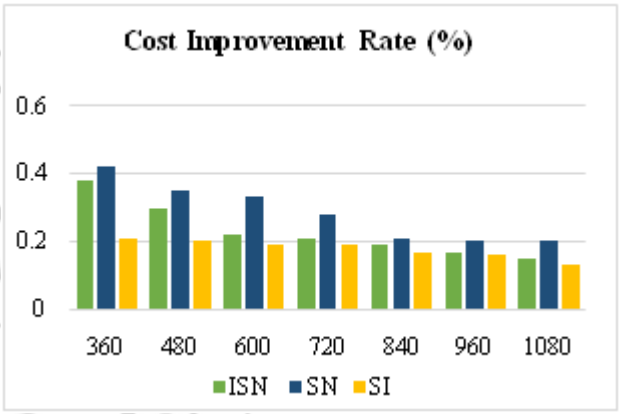

Figure 5 (a): Comparison of Algorithms (Cost Improvement)

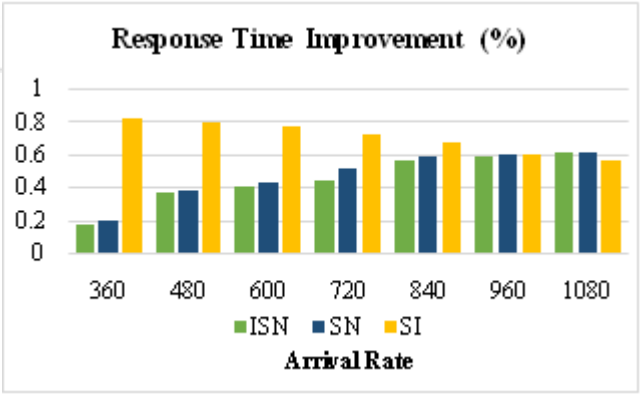

Figure 5 (b): Comparison of algorithms (Response time improvement)

5. Security Mechanisms in Cloud Computing

5.1 Encryption and Public Key Infrastructure

Data, by default is coded in plaintext. When transmitted over a network, plaintext is vulnerable to unauthorized and malicious access. Encryption technology commonly relies on ciphers to encrypt plaintext, which is then referred to as cipher text [22]. Access to cipher text only reveals certain information such as message length and date. It does not reveal the message itself. Encryption mechanism can help counter traffic eavesdropping, malicious changes to data and security threats. Malicious agents that attempt to capture data are unable to decrypt messages if they do not have the encryption key.

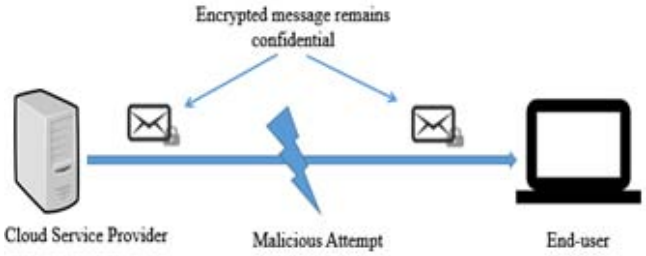

Figure 6: Encryption in the cloud

Symmetric encryption is a technique in which the same key is used to encrypt and decrypt a message. However, it is not as effective as Asymmetric encryption, since in the event of the attacker gaining control of the key, they can decrypt the message. Asymmetric encryption uses a public key available to all to encrypt a message but this message can only be decrypted via a private key known only to the recipient.

The Public Key Infrastructure (PKI) is a system of protocols that enable large scale systems to use public key cryptography. PKIs rely on digital certificates which are digitally signed data structures that bind public keys to certificate owners and related information such as validity periods. Digital Certificates are signed by a third-party Certificate Authority (CA). The PKI role that assures valid and correct registration is called registration authority (RA). An RA is responsible for accepting requests for digital certificates [23] and authenticating the entity making the request.

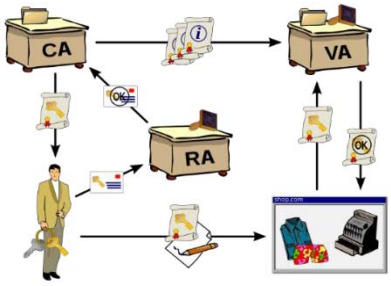

Figure 7: Public Key Infrastructure [24]

5.2 Single Sign-On (SSO)

Managing the authentication and authorization information for a cloud service consumer across multiple applications can be a challenge, especially if numerous resources are to be invoked as part of one runtime activity. The Single Sign-On (SSO) enables the cloud user to be authenticated by a security broker, which establishes a security context that persists while the user access multiple cloud applications. In other case, the user would have to authenticate themselves on every subsequent request.

The single sign-on is a way to access the multiple, related, but independent software system in such a way that user logs in a system and gains the access to all the system without being prompted to re-login in each application. There are several advantages [25] of single sign-on. These are:

- It increases the productivity of the organization. The user is not mired by multiple logins and is not required to remember username and password for each application.
- It simplifies the IT administration by reducing the number of usernames/passwords that should be managed.
- It increases the security of the system.

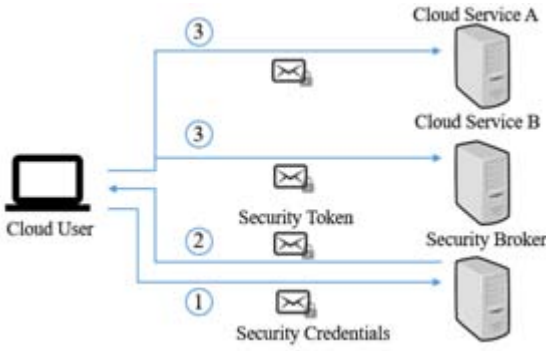

Figure 8: A user provides credentials to security broker (1). The broker responds with an authentication token upon successful authentication (2) that is used to authenticate user across cloud services A & B.

5.3. Hardened Virtual Server Images

A virtual server is created from a virtual image (or a virtual machine image). Hardening of a virtual images aims at stripping off unnecessary software from an operating system to limit vulnerabilities in a system [26]. Redundant programs are removed, disabling root accounts, server ports and guest access are all steps in hardening a virtual image. The hardened virtual image is significantly secure than the original virtual image. Hardening of a virtual image helps counter insufficient authorization, denial of service attacks (DoS) etc.

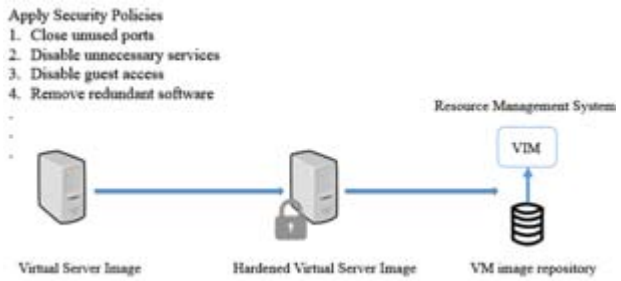

Figure 9: A cloud provider applies its policies to harden the virtual server image. The hardened template is saved in the image repository as part of resource management system.

6. Conclusion

Power is an extremely scarce resource and hence effective power saving methods must be adopted in existing and future cloud system designs. In the experiments conducted by Chiang et al. [10], the SI policy proved to achieve greater cost effectiveness when there is a lower startup cost. It also improves response time in a low arrival rate situation significantly. But, in a converse situation other policies obtain more benefits. Positioning future datacenters underwater can also prove to significantly reduce power consumption due to the natural cooling offered by the water and they also reduce data latency.

Guaranteeing user and data privacy is the primary concern of the cloud. Authentication methods must be secure and difficult to break into. Authentication methods define the authenticity of the user. PKI is by far the most common authentication method with different algorithms such as AES, MD5 proving to be difficult to crack by an attacker. To ensure a secure cloud environment, cloud providers must make sure their servers are secure and immune to attacks and need to address issues pertaining to network security, data integrity, authorization, confidentiality and various other factors.

To achieve a secure cloud computing environment, security threats must be studied and addressed accordingly.

7. Acknowledgement

I would like to take this opportunity to express my profound gratitude and deep regard to my guide, Mrs. Vina M. Lomte, H.O.D, Computer Engineering, RMDSSOE and the Academic Staff of the Department for their exemplary guidance, valuable feedback and constant encouragement and support. I would also like to thank Yi-Ju Chiang, Yen-Chieh Ouyangand Ching-Hsien Hsufor their excellent work.

References

[1] Peter Mell, Timothy Grance "The NIST Definition of Cloud Computing" *Recommendations of the National Institute of Standards and Technology*NIST Special Publication, U.S. Dept. of commerce 800-145 Sept. 2011

[2] "Amazon Web Services", https://aws.amazon.com/

[3] C.R. Paul, "Introduction to Electromagnetic Compatibility,"*New York: Wiley-Interscience*s, 1992, pp. 402-428.

[4] Bill Snyder (2010, December 31) "Server Virtualization has stalled, despite the hype", Retrieved from http://www.infoworld.com/article/2624771/server-virtualization/server-virtualization-has-stalled--despite-the-hype.html

[5] "The Growth of Cloud Computing", http://www.contegix.com/the-growth-of-cloud-computing/

[6] Y. C. Lee, and A. Y. Zomaya, "Energy efficient utilization of resources in cloud computing systems," *The Journal of Supercomputing*, vol. 60, no. 2, pp. 268-280, 2012.

[7] A. Beloglazov et al., "A taxonomy and survey of energy-efficient data centers and cloud computing systems," *Advances in Computers*, vol. 82, pp. 47–111, 2011.

[8] N. Ani Brown Mary et al, "An Extensive Survey on QoS in Cloud computing"*(IJCSIT) International Journal of Computer Science and Information Technologies*, Vol. 5 (1), 2014, 1-5 ISSN: 0975-9646

[9] M. Yadin, M, and P. Naor, "Queueing systems with a removable service station," *Operations research quarterly 14*, pp. 393-405, 1963.

[10] Chiang et al., "An Efficient Green Control Algorithm in Cloud Computing for Cost Optimization", *IEEE TRANSACTIONS ON CLOUD COMPUTING*, TCCSI-2014-03-0116

[11] Hayato Huseman (2015, July 23) "There is no such thing as the cloud", Retrieved from: http://pocketnow.com/2015/07/23/no-such-thing-as-the-cloud

[12] Amlan Deep Borah et al."Power Saving Strategies in Green Cloud Computing Systems", *International Journal of Grid Distribution Computing* Vol.8, No.1 (2015), pp.299-306

[13] W. Huang et al., "An Energy Efficient Virtual Machine Placement Algorithm with Balanced Resource Utilization," *Seventh International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS)*, pp. 313-319, 2013.

[14] Athima Chansanchai (2016, February 1) "Microsoft research project puts cloud in the ocean for the first time." Retrieved from: http://news.microsoft.com/features/microsoft-research-project-puts-cloud-in-ocean-for-the-first-time/

[15] Cisco (2011) "Power Management in the Cisco Unified Computing System: An Integrated Approach" Retrieved from: http://www.cisco.com/c/en/us/solutions/collateral/data-center-virtualization/unified-computing/white_paper_c11-627731.html

[16] (2013, July 17) "Facts and Stats of World"s largest data centers" Retrieved from: http://storageservers.wordpress.com/2013/07/17/facts-and-stats-of-worlds-largest-data-centers/

[17] Y. Deng et al., "M/M/1 queueing system with delayed controlled vacation," *OR Transactions*, vol. 3, pp. 17-30, 1999.

[18] K. H. Wang and H. M. Huang, "Optimal control of an M/Ek/1 queueing system with a removable service station," *Journal of the operational research society*, vol. 46, pp.1014 -1022, 1995.

[19] M. Zhang, and Z. Hou, "M/G/1 queue with single working vacation," *Journal of Applied Mathematics and Computing*, vol. 39, no. 1-2, 221-234, 2012.

[20] Project Natick http://www.projectnatick.com/

[21] Jacob Nganyi et al. (2010, April 7) "Human Settlements on the Coast"Retrieved from: http://www.oceansatlas.org/servlet/CDSServlet?status=ND0xODc3JjY9ZW4mMzM9KiYzNz1rb3M~

[22] Thomas Erl et al. "Cloud Security Mechanisms" in *Cloud Computing Concepts, Technology & Architecture*, 3rd ed. Noida, India, Pearson 2016, 229-255

[23] Techotopia, (2016, February 1) "An overview of Public Key Infrastructure" Retrieved from: http://www.techotopia.com/index.php/An_Overview_of_Public_Key_Infrastructures_(PKI)

[24] Image by Chris 論 - [1] and OpenCliparts.org, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=2501151

[25] Paras Tiwari, Shashidhar Joshi "Single Sign-on with One Time Password" *Internet, 2009. AH-ICI 2009. First Asian Himalayas International Conference on*, pages 1-4 10.1109/AHICI.2009.5340290

[26] "Hardened Virtual Server Image" http://cloudpatterns.org/mechanisms/hardened_virtual_server_image