Is the Cloud the Future of Computing?

Joseph M. Kizza and Li Yang
Department of Computer Science and Engineering
The University of Tennessee-Chattanooga, Chattanooga, Tennessee

Abstract

Cloud computing as a technology is difficult to define because it is evolving without a clear start point and no clear prediction of its future course. Even though this is the case, one can say that it is a continuous evolution of a computer network technology going beyond the client-server technology. It is a technology extending the realms of a computer network creating an environment that offers scalability, better utilization of hardware, on-demand applications and storage, and lower costs over the long run through the creation of virtual servers cloned from existing instances each offering near instantaneous increase in performance, allowing companies to react quickly and dynamically to emerging demands. The "cloud" or "cloud solution", as the technology is commonly referred to, can either be hosted onsite by the company or off-site such as Microsoft's SkyDrive and Samsung's S-Cloud.

The cloud technology seems to be in flux; hence it may be one of the foundations of the next generation of computing. Keep watching! It next a few years, a grid of a few cloud infrastructures may provide computing for millions of users. This is a broader view of cloud computing. Cloud computing technology consists of and rests on a number of sound, fundamental and proven technologies including virtualization, service oriented architectures, distributed computing, grid computing, broadband networks, software as a service, browser as a platform, free and open source software, autonomic systems, web application frameworks and service level agreements. Based on these fundamental and sound computing principles, one wonders whether cloud computing is the next trajectory of computing. This chapter will discuss this in depth and also look at the security issues involved.

1. Introduction

Cloud computing as a technology, in its present form, is difficult to define because it is evolving without a clear start point and no clear prediction of its future course is known yet. However, one can say that cloud computing has gone beyond the client-server paradigm in networking environment which offers scalability, increased utilization of hardware, on-demand software applications and storage. Cloud computing lowers cost of operation over the long run through employing virtual servers which lead to instantaneous increased performance and fast response to any emerging hardware, software or service demands. With the current trends in cloud technology, it may be that in the next few years, a grid of a few cloud infrastructures may provide computing for millions of users.

Cloud computing technology consists of and rests on a number of sound, fundamental and proven fundamental technologies including virtualization, service oriented architectures, distributed computing, grid computing, broadband networks, software as a service, browser as a platform, free and open source software, autonomic systems, web application frameworks and service level agreements (Mell, 2011). We will discuss cloud computing based on these technologies.

First let us start by giving a broader but specific view of the technology, what it is composed of and how it works. According to NIST(Mell, 2011), cloud computing is a model for enabling

ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources like networks, servers, storage, applications and services that can be rapidly provisioned and released with minimal management effort or service provider interaction. So for the remainder of this chapter, we are going to focus on this model of computing and discuss its benefits and security concerns.

2 Historical Development of the Cloud Infrastructure

Traditionally data center computing models were mainly based on a client-server model architecture and design relying firmly on a three-tier architecture design that included access, distribution and core switches connecting relatively few clients and meeting limited client needs compared to today's cloud services models. Each server was dedicated to either a single or limited applications and had IP addresses and media access control addresses. This static nature of the application environment worked well and lent itself to manual processes for server deployment or redeployment. According to both Jim Metzler and Steve Taylor of Network World (Metzler, 2011), they primarily used a spanning tree protocol to avoid loops. Recent dramatic advances in virtualization technology, distributed computing, rapid improvements and access to high-speed Internet have all had dramatic influences on the current models of computing and data center. From services on demand to unprecedented elasticity in resource acquisition, users now have an array of choices at hand on demand and in quantities of choice. The services are fully managed by the provider, with the user as a consumer. Let us briefly look at those characteristics that have come to define cloud computing as a technology (Mell, 2011).

Ubiquitous network access

The recent ubiquitous access to computer networks and services attribute to advances and use of high speed Internet and virtualization technology. Advances and development in these technologies have increased options the repertoire of computing services a customer can select from. With more option also came the high specialization and quality of services that a customer can expect.

Measured service

The increase in the repertoire of services available to users has been enhanced by cloud services' elasticity, flexibility, on demand capabilities thus allowing for these services to be metered. The concept of metered services allows customers to get what they want in the required amounts at the time they want the service. One of the most popular characteristics of cloud computing technology is measured or metered service for most, if not all, of the cloud services including storage, processing, bandwidth and active user accounts. This *pick-what-you-can-afford-to-pay-for* principle based on metering results in an automatic control and optimization of cloud technology resource use based on the type of service and these statistics can be reported as needed thus providing transparency for both the provider and consumer.

On-demand self-service

Traditionally, acquisition of computing services demanded perpetual ownership of software or computing hardware and sustainable technical support to help with computing services. Those models are phasing out when we have cloud computing as a flexible model where consumers of computing services are no longer restricted to rigid traditional models of ownership or boxed services. Now, a consumer is able to not only automatically provision any computing services

and capabilities as needed but also to determine the time and how long to use the provisioned services.

Rapid elasticity

The ability to resize and dynamically scale the virtualized computing resources at hand such as servers, processors, operating systems and others to meet the customer's on-demand needs is referred to as computing service elasticity. To meet elasticity demands on computing resources, the provider must make sure that there are abundant resources at hand that to ensure that end-users' requests are continually and promptly met. Amazon's EC2 is a good example of a web service interface that allows the customer to obtain and configure capacity with minimal effort.

Resource pooling

As noted in the NIST report, the provider's computing resources are pooled to serve multiple consumers using a multi-tenant model, with different physical and virtual resources dynamically assigned and reassigned according to consumer demand. These fluctuating and unpredictable customer demands are a result of new cloud computing flexibility, access and ease of use.

There are other characteristics common to cloud computing beyond the five we have discussed above. Among these are (Mell, 2011):

- Massive scale – that the cloud offers the resources at a massive scale on demand.
- Virtualization – in fact this is the linchpin of the cloud technology. The cloud is possible because of virtualization of the fundamental functionalities of the physical machine.
- Free software – or near free software as needed from the cloud.
- Autonomic computing – in a sense that you scale computing resources at a time you want them on the fly.
- Multi-tenancy – because of cloud's massive scale and easy access of those resources, cloud computing can accommodate a large number of users at a time.

3 Cloud Computing Service Models

Infrastructure as a Service (IaaS). Cloud computing offers flexibility and autonomy that allow customers to manage and control system resources via a web-based virtual server instance API. Customers are able to start, stop, access, and configure operating systems, applications, storage and other fundamental computing resources without interacting with the underlying physical cloud infrastructure.

Platform as a Service (PaaS). This is a set of software and product development tools hosted on the provider's infrastructure and accessible to the customer via a web-based virtual server instance API. Through this instance, the customer can create applications on the provider's platform over the Internet. Accessing the platform via the web-based virtual instance API protects the resources because the customer cannot manage or control the underlying physical cloud infrastructure including network, servers, operating systems, or storage.

Software as a Service (SaaS). Ever since the beginning of computing software, over the years, the key issue that has driven software development has been the issue of the cost of software. Trying to control the cost of software has resulted into software going through several models. The first model was the home developed software where software users developed their own

software based on their needs and they owned everything and were responsible for updates and management of it. The second model, the traditional software model was based on packaged software where the customer acquired more general purpose software from the provider with a license held by the provider and the provider being responsible for the updates while the customer is responsible for its management. However, sometimes, software producers provide additional support services, the so called premium support, usually for additional fees. Model three was the Open Source model led by a free software movement starting around the late 80s. By the late 1980, free software turned into open source with the creation of the Open Source Initiative (OSI). Under the name of "open source" philosophy, some for-profit "free software" started to change the model from purely free software to some form of payment in order to support updates of the software. The open source software model transformed the cost of software remarkably. Model Four consisted of Software Outsourcing.

The outsourcing model was in response to the escalating cost of software associated with software management. The component of software management in the overall cost of software was slowly surpassing all the costs of other components of software including licensing and updates. In model four, the software producer takes on the responsibility of the management of that software because software is still licensed from the software company on a perpetual basis.

Software model five is Software as a Service (SaaS). Under this model, there is a different way of purchasing. Under SaaS, there is the elimination of the upfront license fee. All software applications are retained by the provider and the customer has access to all applications of choice from the provider via various client devices through either a thin client interface, such as a web browser, a web portal or a virtual server instance API. The cloud user's responsibilities and actual activities in the use of and operations of the requested cloud services is limited to user-specific application configuration settings, leaving the management and control of the underlying cloud infrastructure including network, servers, operating systems, storage, or even individual application capabilities to the cloud provider.

4 Cloud Computing Deployment Models

There are four cloud deployment models: public, private, hybrid, and community (Mell, 2011).

Public clouds- The public clouds provides access to computing resources for the general public over the Internet allowing customers to self-provision resources typically via a web service interface on a pay-as-you-go basis. One of the benefits of public clouds is to offer large pools of scalable resources on a temporary basis without the need for capital investment in infrastructure by the user.

Private cloud - Unlike public clouds, private clouds give users immediate access to computing resources hosted within an organization's infrastructure and premises. Users, who are usually in some form of a relationship with the cloud owner, choose and scale collections of resources drawn from the private cloud, typically via web service interface, just as with a public cloud. Also the private cloud is deployed within and uses the organization's existing resources and is always behind the organization's firewall subject to the organization's physical, electronic, and procedural security measures. Security concerns are addressed through secure-access virtual private network (VPN) or by the physical location within the client's firewall system. In this case, therefore, private clouds offer a higher degree of security.

Hybrid cloud - A hybrid cloud combines the computing resources of both the public and private

clouds, which helps businesses to take advantage of secured applications and data hosting on a private cloud, while still enjoying cost benefits by keeping shared data and applications on the public cloud. This model is also used for handling cloud bursting, which refers to a scenario where the existing private cloud infrastructure is not able to handle load spikes and requires a fallback option to support the load. Hence, the cloud migrates workloads between public and private hosting without any inconvenience to the users. Many PaaS deployments expose their APIs, which can be further integrated with internal applications or applications hosted on a private cloud, while still maintaining the security aspects. Microsoft Azure and Force.com are two examples of this model.

Community cloud – A community cloud is shared by several organizations with the same policy and compliance considerations. This helps to further reduce costs as compared to a private cloud, as it is shared by larger group. Some state-level government departments requiring access to the same data relating to the local population or information related to infrastructure, such as hospitals, roads, electrical stations, etc., can utilize a community cloud to manage applications and data.

5 Benefits of Cloud Computing

Cloud computing as a model of computing is very exciting and has tremendous benefits for those who dare to use it. It is not only exciting when you come to learn it, but it also has an array of benefits including but not limited to leveraging on a massive scale, homogeneity, virtualization, low cost software, service orientation, and advanced security technologies (Mell, 2011).

Reduced Cost – The biggest benefit from all cloud computing benefits to a company perhaps lies in cost savings. Whether it is a small, medium or large scale manufacturing business there are essential cost benefits in using a cloud model for most of the company's computing needs. The biggest issue here is the fact that cloud computing is operated remotely off company premises except a few devices needed for accessing the cloud resources via a web portal. This means that company personnel can do the same amount of work on fewer computers by having higher utilization, save on not housing data centers on premises, save on personnel for running the data center, save on expenses that would normally be essential for running a data center on the premises. There are documentary evidences to support these views from industry experts. In the words of Greg Papadopoulos, the CTO from Sun Microsystems (Farber, 2008), hosting providers bring 'brutal efficiency' for utilization, power, security, service levels, and idea-to-deploy time. And there are savings on power consumption since there are few computers on premises. Currently, servers are used at only 15% of their capacity in many companies and 80% of enterprise software expenditure is on installation and maintenance of software.

Automatic Updates- Our economy is now an online economy because most of, if not all businesses, are now online and depend on software applications for day to day services. Software is continuously changing and as business functionalities change, software need to be changed or updated. The cost of software updates and management has always been on the rise, usually surpassing the cost of new software. For companies to stay competitive and in many cases afloat, they must be consistently updating and changing software. The business of software updates and software management and licensing is a big drain on company resources. So having automatic updates and management from the cloud provider can be a great relief to any company. But updates are not limited to only software. Also not worrying about hardware updates is cost effective for companies.

Green Benefits of Cloud computing - Although cloud computing energy consumption has seen a vigorous debate and this debate is continuing, pitting those claiming that cloud computing is gobbling up resources as large cloud and social networking sites need daily megawatts of power to feed insatiable computing needs and those who claim that the computing model is indeed saving power from millions of servers left idling daily and consuming more power. We will discuss this more in the coming sections. For now, we think that there are indeed savings in power consumption by cloud computing.

Remote Access- With a web portal access to the cloud, company employees may be able to work while they are on the road, home or in the office. This is of great benefit to the company so that there is no down time because somebody is not in the office.

Disaster Relief- Many companies live in constant fear disasters occurring when they have company vital data stored on premises. No one likes to be a victim of large-scale catastrophes such as devastating hurricanes, earthquakes, and fires and of course terrorist attacks. Such misfortunes can create havoc to the companies' vital data and disrupt operations even if there were limited physical damage. Additionally, there are smaller disasters like computer crashes and power outages that can also wreak havoc on a company's vital data. While this is possible, there are many companies, especially small ones that may not even have any disaster recovery plan and some who have it may not be able to execute it effectively. This fear can be overcome with investments in cloud technology. Company's vital back up data can be safely stored on secure data centers on the cloud instead of in the company's server room.

Self-service provisioning - Cloud computing allows users to deploy their own virtual sets of computing resources like servers, network, storage, and others, as needed without the delays, competency and complications typically involved in physical resource acquisition, installation and management. The cloud owners, irrespective of their physical location, not only can they provide all the computing resources your organization needs but also have the necessary capacity needed to monitor, manage and respond to the organization's daily and hour by hour infrastructure, software and platform requirements.

Scalability - Because of the minute by minute monitoring capability of cloud computing of an organization's computing needs and the ability to increase or reduce the required resources as the demand increases or decreases, cloud computing offer the best infrastructure, platform and software scalability that cannot be matched in any owned computing facility.

Reliability and fault-tolerance - Because the cloud provider, with qualified professionals and experience, monitors the computing requirements of a client company and can easily scale to demand, cloud computing offers a high degree of reliability and fault-tolerance.

Ease of Use - To attract more customers, cloud provider has and must make the user interface friendly so that customers can scale into the cloud with the least effort.

Skills and Proficiency – Some of the most sought after assets from a cloud provider are professionalism and a vast skill set provided to the customers. Companies, especially small ones, would pay a high price to get an employee with the skills set, efficiency, proficiency and experience found with cloud center staff.

Response Time- Depending on the bandwidth at the company web portal, cloud computing services normally have speed because the computing resources provided are modern and powerful to be able to accommodate large number of users.

Mobility – Because of web portal interface to the Cloud, cloud computing essentially is a mobile computing platform, allowing the users to access their applications.

Increased Storage – Storage is a main function from cloud computing. Because of this, it is cheap and readily scalable to need.

Other Benefits - Other benefits beyond those we descussed above include, providing a high quality of service (QoS), providing a high quality, well-defined and stable industry standard API and on demand availability of computing resources based on "at hand" financial contraints.

Security – We are going to discuss this more in the coming section, but cloud computing, because of its individual virtual machines created per use, security provision has already been built in. In addition to these built in provisions due to virtualization, the Cloud model also offers a strong authentication regime at the browser interface gateway, a security mechanism that is individually and quickly set up and torn down as needed and a strong validation and verification scheme that is expensive to deploy at an individual client-server model.

6 Security, Reliability, Availability and Compliance Issues of Cloud Computing

The cloud computing model as we know it today did not start overnight. The process has taken years moving through seven software models beginning with in-house software, licensed software normally referred as the traditional model, open source, outsourcing, hybrid, software as a service and finally the Internet model, the last two being part of the cloud computing model. When one carefully examines the cloud servicing model, one does not fail to notice the backward compatibilities or the carryovers of many of the attributes that characterized software through all the models. While this brings the benefits of each one of those software models, but also many, if not all of the software complexity and security issues in those models were carried over into the cloud computing model. Because of this, our first thought was to discuss the security issues in the cloud computing model through the prism of these models. It is tempting but we are going to follow a different path while keeping the reader rooted into the different software models. Security is and continues to be a top issue in the cloud computing model. The other three related issues are performance, compliance and availability. We will discuss all four in this section but since security is the number one issue, we will address it first.

We want to start the discussion of cloud computing security by paraphrasing Greg Papadopoulos, CTO of Sun Microsystems who said that cloud users normally "trust" cloud service providers with their data like they trust banks with their money. This means that they expect the three issues of security, availability and performance to be of little concern to them as they are with their banks (Farber, 2008). To give a fair discussion of the security of anything, one has to focus on two things that are the actors and their roles in the process you are interested in securing and the application or data in play. The application or data is thought of in relation to the temporal state it is in. For example the states of data are either in motion between the remote hosts and the service provider's hypervisors and servers or in the static state when it is stored at remote hosts, usually on the customer's premises or in the service provider's servers. The kind of security needed in either one of these two states is different.

6.1 Delegated Security Responsibilities in the Cloud

In the cloud computing model, the main players are the cloud provider, the customer who is the data owner and who seeks cloud services from the cloud provider and the user who may or may not be the owner of the data stored in the cloud. The security responsibilities and expectations of

each one of these players are different. The first two players have delegated responsibilities to all who work on their behalf. To fully understand these delegated responsibilities assigned to each one of these, we need to look at first the marginal security concerns resulting from the peripheral system access control that always result in easiest breach of security for any system, usually through compromising user accounts via weak passwords. This problem is broad affecting both local and outsourced cloud solutions. Addressing this and all other administrative and use security concern requires companies offering and using cloud solutions to design an access control regime that covers and requires every user, local or remote, to abide by these access policies including the peripheral ones like the generation and storage of user passwords. Additionally, employees need to be informed of the danger of picking easy passwords and to understand the danger of writing a password down or divulging a password to anyone. Access control administration is so important that cloud providers spend large amounts of resources to design a strong assess control regimes.

6.2 Security of Data and Applications in the Cloud

Security is arguably the most relevant concern preventing players' entry in the cloud. Threats posing to data, applications and users in the cloud are determined by a large number of technologies it comprises. The most relevant threats for the cloud have recently been shown in (Archer, 2010).

- Threat 1: *Abuse and Nefarious Use of cloud computing*: IaaS providers bring the illusion of unlimited compute, network, and storage capacity. However, IaaS offerings have hosted the botnets, Trojan horses, and software exploits.
- Threat 2: *Insecure Interfaces and APIs*: Cloud Computing providers expose a set of software interfaces or APIs that customers use to manage and interact with cloud services. Provisioning, management, orchestration, and monitoring are all performed using these interfaces. The security and availability of general cloud services is dependent upon the security of these basic APIs.
- Threat 3: *Malicious Insiders*: this is amplified in the cloud by "the convergence of IT services and customers under a single management domain, combined with a general lack of transparency into provider process and procedure".
- Threat 4: *Shared Technology Issues*: IaaS vendors deliver their services in a scalable way by sharing infrastructure. Monitor environment for unauthorized changes/activity.
- Threat 5: *Data Loss or Leakage*: This threat increases in the cloud, due to the architectural or operational characteristics of the cloud environment (e.g. insecure APIs, shared environment, etc.).
- Threat 6: *Account or Service Hijacking*: phishing, fraud, and exploitation are well known issues in IT. The cloud adds a new dimension to this threat: "if an attacker gains access to your credentials, they can eavesdrop on your activities and transactions, manipulate data, return falsified information, and redirect your clients to illegitimate sites. Your account or service instances may become a new base for the attacker".
- Threat 7: *Unknown Security Profile*: the reduction of cost of ownership induced by the cloud also resulted in more complex analysis of a company's security posture. More tenants imply increased complexity in detecting who and how the cloud infrastructure is used.

We will then discuss the security of data and applications in the cloud. To do this we need to focus first on the security and role of the hypervisor and then the servers on which user services are based. A hypervisor is also called virtual machine manager (VMM), which is one of many hardware virtualization techniques allowing multiple operating systems to run concurrently on a host computer. The hypervisor is piggybacked on a kernel program, itself running on the core physical machine running as the physical server. The hypervisor presents to the guest operating systems a virtual operating platform and manages the execution of the guest operating systems. Multiple instances of a variety of operating systems may share the virtualized hardware resources. Hypervisors are very commonly installed on server hardware, with the function of running guest operating systems, that themselves act as servers. The security of the hypervisor therefore involves the security of the underlying kernel program and the underlying physical machine, the physical server and the individual virtual operating systems and their anchoring virtual machines.

6.2.1 Hypervisor Security

The key feature of the cloud computing model is the concept of virtualization. It is virtualization that gives the cloud the near instant scalability and versatility that makes cloud computing so desirable a computing solution by companies and individuals. The core of virtualization in cloud computing is the easy process of minting of virtual machines on demand by the hypervisor. The hypervisor allocates resources to each virtual machine it creates and it also handles the deletion of virtual machines. Since each virtual machine is initiated by an instance, the hypervisor is a bi-directional conduit into and out of every virtual machine. The compromise of either, therefore, creates a danger to the other. However, most hypervisors are constructed in such a way that there is a separation between the environments of the sandboxes (the virtual machines) and the hypervisor. There is just one hypervisor, which services all virtual sandboxes, each running a guest operating system. The hypervisor runs as part of the native monolithic operating system, side-by-side with the device drivers, file system and network stack, completely in kernel space. So, one of the biggest security concerns with a hypervisor is the establishment of covert channels by an intruder. According to the Trusted Computer Security Evaluation Criteria, TCSEC, a covert channel is created by a sender process that modulates some condition (such as free space, availability of some service, wait time to execute) that can be detected by a receiving process. If an intruder succeeds in establishing a covert channel, either by modifying file contents or through timing, it is possible for information to leak from one virtual machine instance to another (Violino, 2010).

Also since the hypervisor is the controller of all virtual machines, it, therefore, becomes the single point of failure in any cloud computing architecture. That is, if an intruder compromises a hypervisor then the intruder has control of all the virtual machines the hypervisor has allocated. This means that the intruder can even create or destroy virtual machines at will. For example, the intruder can perform a denial of service attack, by bringing down the hypervisor which then brings down all virtual machines running on top of the hypervisor.

The processes of securing virtual hosts differ greatly from processes used to secure their physical counterparts. Securing virtual entities like a hypervisor, virtual operating systems and corresponding virtual machines is more complex. To understand hypervisor security, let us first discuss the environment in which the hypervisor works. Recall that a hypervisor is part of a Virtual Computer System (VCS). In his 1973 thesis in the Division of Engineering and Applied Physics, Harvard University, Robert P. Goldberg defines a virtual computer system as a

hardware-software duplicate of a real existing computer system in which a statistically dominant subset of the virtual processor's instructions execute directly on the host processor in native mode. He also gives two parts to this definition, the environment and implementation (Goldberg, 1973).

Environment – That the virtual computer system must simulate a real existing computer system. Programs and operating systems which run on the real system must run on the virtual system with identical effect. Since the simulated machine may run at a different speed from the real one, timing dependent processor and I/O code may not perform exactly as intended.

Implementation- Most instructions being executed must be processed directly by the host CPU without recourse to instruction by instruction interpretation. This guarantees that the virtual machine will run on the host with relative efficiency. It also compels the virtual machine to be similar or identical to the host, and forbids tampering with the control store to add an entirely new order code.

In the environment of virtual machines, a hypervisor is needed to control all the sandboxes (virtual machines). Generally in practice, the underlying architecture of the hypervisor determines if there is a desired true separation between the sandboxes. Robert P. Goldberg classifies two types of hypervisor (Goldberg, 1973):

Type-1 (or *native*, *bare metal*) hypervisors run directly on the host's hardware to control the hardware and to manage guest operating systems. See figure 1 below. All guest operating systems then run on a level above the hypervisor. This model represents the classic implementation of virtual machine architectures. Modern hypervisors based on this model include Citrix XenServer, VMware ESX/ESXi, and Microsoft Hyper-V. The most common commercial hypervisors are based on a monolithic architecture below.

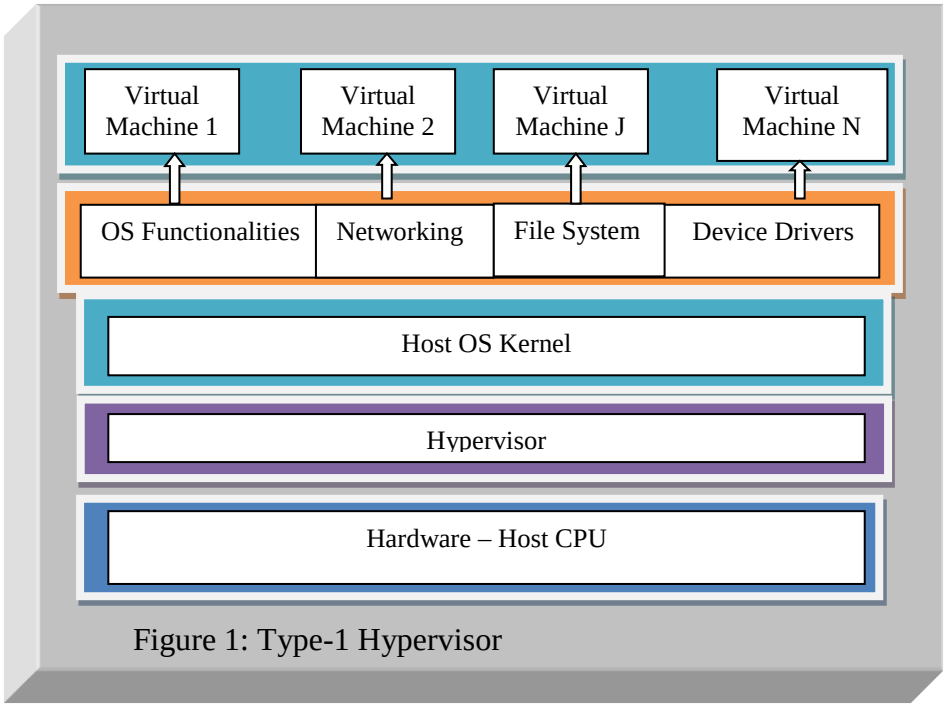

Figure 1: Type-1 Hypervisor

The underlying hypervisor services all virtual sandboxes, each running a guest operating system. The hypervisor runs as part of the native monolithic operating system, side-by-side with the device drivers, file system and network stack, completely in kernel space.

Type-2 (or *hosted*) hypervisors run just above a host operating system kernel such as Linux, Windows and others as in figure 2 below. With the hypervisor layer as a distinct second software level, guest operating systems run at the third level above the hardware. The host operating system has direct access to the server's hardware like host CPU, memory and I/O devices and is responsible for managing basic OS services. The Hypervisor creates virtual machine environments and coordinates calls to CPU, memory, disk, network, and other resources through the host OS. Modern hypervisors based on this model include KVM and VirtualBox.

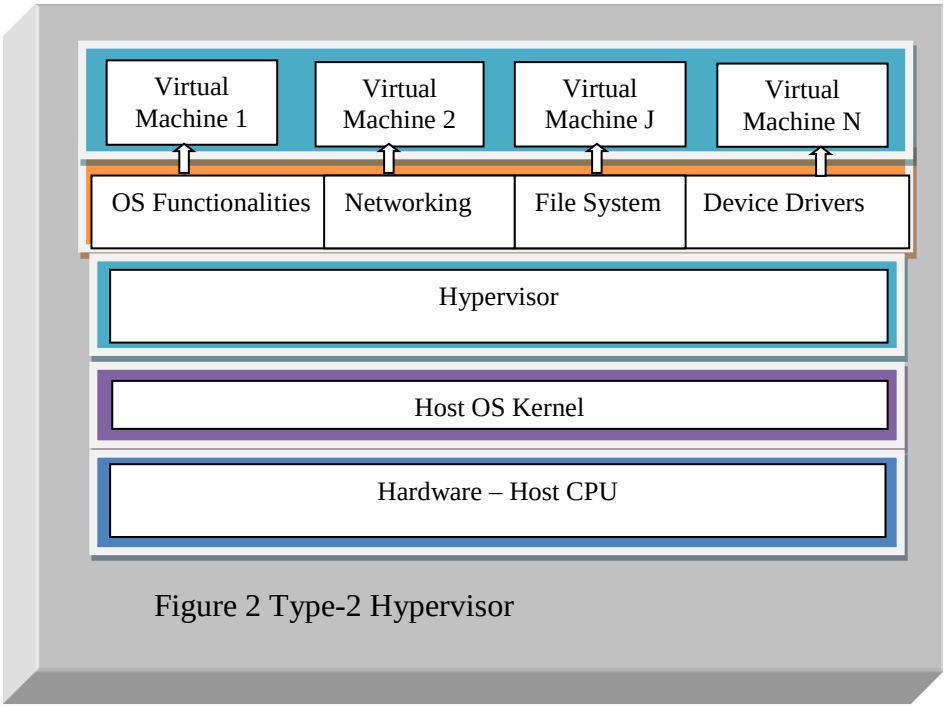

Figure 2 Type-2 Hypervisor

The discussion so far highlights the central role of the hypervisor in the operations of virtual machine systems and it points to its central role in securing all virtual machine systems. Before we look at what can be done to secure it, let us ask ourselves what security breaches can happen to the hypervisor. Some malicious software such as rootkit masquerade themselves as hypervisors in self-installation phases.

Neil MacDonald, Vice President and a Gartner Fellow (MacDonald, 2011) reported his observation , about hypervisor and the vulnerabilities associated with it in his blog titled as “Yes, Hypervisors Are Vulnerable”. His observation is summarized below:

- The virtualization platform (hypervisor/VMM) is software written by human beings and will contain vulnerabilities. Microsoft, VMware, Citrix, and other, all of them will and have had vulnerabilities.
- Some of these vulnerabilities will result in a breakdown in isolation that the virtualization platform was supposed to enforce.
- Bad guys will target this layer with attacks. The benefits of a compromise of this

layer are simply too great.

- While there have been a few disclosed attacks, it is just a matter of time before a widespread publicly disclosed enterprise breach is tied back to a hypervisor vulnerability.

There have been a growing number of virtualization vulnerabilities. Published papers have so far shown that the security of hypervisors can be undermined. As far back as 2006, Samuel T. King, Peter M. Chen, Yi-Min Wang, Chad Verbowski, Helen J. Wang and Jacob R. Lorch demonstrate in their paper "SubVirt: Implementing malware with virtual machines", the use of type of malware, which called a virtual-machine based rootkit (VMBR), installing a virtual-machine monitor underneath an existing operating system and hoists the original operating system into a virtual machine.

In their study, the authors demonstrated a malware program that started to act as its own hypervisor under Windows. We know that the hypervisor layer of virtualization, playing the core role in the virtualization process is very vulnerable to hacking because this is the weakest link in the data centre. Therefore attacks on hypervisor are on the rise. Data from the IBM X-Force *2010 Mid-Year Trend and Risk Report* show that every year since 2005, vulnerabilities in virtualization server products, the hypervisors, have overshadowed those in workstation products, an indication of the hackers interest in the hypervisors. The report further shows that 35% of the server virtualization vulnerabilities are vulnerabilities that allow an attacker to "escape" from a guest virtual machine to affect other virtual machines, or the hypervisor itself. Note that the hypervisor in *type-1* environment is granted CPU privilege to access all system I/O resources and memory. This makes it a security threat to the whole cloud infrastructure. Just a single vulnerability in the hypervisor itself could result in a hacker gaining access to the entire system, including all the guest operating systems. Because malware run below the entire operating system, there is a growing threat of hackers using malware and rootkits to install themselves as a hypervisor below the operating system thus making them more difficult to detect. In *type-2* hypervisor configuration, figure 2, the microkernel architecture is designed specifically to guarantee a robust separation of application partitions. This architecture puts the complex virtualisation program in user space, thus every guest operating system uses its own instantiation of the virtualization program. In this case, therefore, there is complete separation between the sandboxes (virtual boxes), thus reducing the risks exhibited in type-1 hypervisors.

An attack, therefore, on *type-2* hypervisors can bring down one virtual box, not more and cannot bring down the cloud infrastructure as is the case in *type-1* hypervisors.

According to Samuel T. King et al, overall, virtual-machine based rootkits are hard to detect and remove because their state cannot be accessed by software running in the target system. Further, VMBRs support general-purpose malicious services by allowing such services to run in a separate operating system that is protected from the target system (King, 2006).

6.2.2 Securing Load Balancers

For every hypervisor, there is a load balancer, used to route traffic to different virtual machines to help spread traffic evenly across available machines. A Load balancer in a hypervisor plays a vital role of ensuring a fair distribution of available load to all virtual machines especially during high traffic and ensuring the full utilization of the cloud infrastructure. Elastic load balancers play a central in the cloud infrastructure along the following lines:

- It listens to all traffic destined for the internal network and distributes incoming traffic across the cloud infrastructure.
- Automatically scales its request handling capacity in response to incoming application traffic.
- It creates and manages security groups associated with each instance and provides additional networking and security options if and when needed.
- It can detect the health of the virtual machines and if it detects unhealthy load-balanced virtual machine, it stops routing traffic to it and spreads the load across the remaining healthy virtual machines.
- It supports the ability to stick user sessions to specific virtual machines.
- It supports SSL termination at the Load Balancer, including offloading SSL decryption from application virtual machines, centralized management of SSL certificates, and encryption to backend virtual machines with optional public key authentication.
- It supports use of both the Internet Protocol version 4 and 6 (IPv4 and IPv6).

Due to the load balancer's ability to listen and process all traffic that is destined to the internal network of the cloud, it is a prime target for attackers. If a load balancer was compromised an attacker could listen to traffic and may compromise secure traffic destined to outside the network. Additionally, if the load balancer is compromised along with a virtual machine - traffic could be directed to an unsecure internal server where further attacks are launched (Hotaling, 2003). Because the load balancer is a single point in the cloud infrastructure, it very vulnerable to denial-of-service (DoS) attacks which lead to disruption of cloud activity.

What is the best way to secure the load balancer from attacks? A load balancer is normally secured through proper configuration and monitoring of the balancer's logs. This is achieved through restriction of access to administration of the balancer itself by configuring the load balancer to only accept administrative access over a specific administrative network. This administrative network should be connected to the administrative only network. Limiting access over the administrator network greatly limits the number of users with access to the load balancer (Kizza, 2013).

6.2.3 Virtual Operating Systems Security

Besides the hypervisor, the virtualization system also hosts virtual servers each running either a guest operating system or another hypervisor. And on the peripheral of the virtual machine system are the consoles and hosts. Through each one of these resources, the virtual machine system can be susceptible to security vulnerabilities. Let us briefly look at these below:

Host security

Through hosts like workstations, user gain access to the virtual machine system, hence to the cloud. Two problems are encountered here:

- Escape-to-hypervisor vulnerabilities -- that allow intruders to penetrate the virtual machine from the host.
- Escape-to-host vulnerabilities – that allow vulnerabilities in the virtual machine to move to the hosts.

Guest machines

Guest machines running guest operating system can also pose a security problem to the cloud. However, as we saw in the previous chapter, vulnerabilities in the guest virtual machines are confined to that machine and they rarely affect other machines in the system.

6.3 Security of Data in Transition - Cloud Security Best Practices

With the vulnerabilities in the cloud we have discussed above, what is the best way to protect the user of the cloud? For a cloud customer, the key areas of concerns are virtualization technology security vulnerabilities that may be encountered during the use of the cloud that may affect the customer, unauthorized access to customer data and other resources stored or implemented in the cloud, whether the cloud provider uses strong enough encryption to safeguard customer data, secure access and use of cloud applications and secure cloud management. Let us next discuss the best practices that try to address some of these concerns.

6.4 Service Level Agreements (SLAs)

A service-level agreement (SLA) is a service contract between the provider of a service and the client defining the level of expected service in terms of security, availability and performance. The Cloud service-level agreements (SLAs) are a series of service contracts between cloud providers and clients to define the level(s) of service based on the types of services sought by the client because the effectiveness of these contracts depend on how well maximized and tailored these services are to the particular needs of each client. For example, the security of services sought by a client may depend on the tier of cloud offering the client is using. To see how involved and intricate these documents can be, take an example of security concerns. For IaaS, the security responsibilities are shared with the provider responsible for physical, environmental, and virtualization security, while the client takes care of the security in applications, operating system, and others. Now if we change the service model to SaaS, the provider is responsible for almost every aspect of security.

6.5 Data Encryption

The moment data leaves your end-point web-cloud access point in your location, it travels via a public network and stored in shared environment – the cloud. In a public or in a shared environment, data can be intercepted and infiltrated by intruders from within and outside the cloud and during transmission from man-in-the-middle crypto-analysis. To prevent these kinds of breaches strong encryption and authentication regimes are needed. Encryption to safeguard any kind of data breaches require a strong access control and authentication to all web-based cloud resource interfaces, encryption of all administrative access to the cloud hypervisor, all access to applications and data.

6.6 Interface and API Security

Most cloud access instances are web-based. A set of APIs to manage and interact with cloud services are typically exposed (provisioning, monitoring, etc.). Most security breaches to stored data originated from Web applications or APIs. Cloud security therefore depends on strong security controls of these basic cloud APIs.

7. Future Challenges

Many of the cloud security issues just reflect traditional web application, networking, and data-hosting problems although we have been relating them with cloud specific element throughout

the chapter. Issues, including phishing, downtime, data loss, password weakness, and compromised hosts running botnets, will remain to be challenges in cloud computing environments. *Scalability* is a paramount problem for securing current clouds at the level of physical infrastructure. Denial of service (DoS) attacks is easier in an environment with a high number of cloud users if not appropriately managed. *VM image management* is an issue since VM images need to be moved from in-house trusted facilities to a cloud provider through unsecured networks. VM encryption techniques are weakened by the fact that VMs are usually large files. *Virtual networks* are also subject to some security concerns such as how to securely and dynamically establish data paths for communicating distant VMs. The cloud leads to a drop in security as the traditional controls such as virtual local-area networks (VLANs) and firewalls prove less effective during the transition to a virtualized environment. Trust management and auditability remain challenging cloud security. The establish zones of trust in the cloud, the virtual machines must be self-defending, effectively moving the perimeter to the virtual machine itself. Enterprise perimeter security only controls the data that resides and transits behind the perimeter. In the cloud computing world, the cloud computing provider is in charge of customer data security and privacy. The assumption of full trust in IaaS providers might not be true. The complex chain of trust introduced by different cloud stakeholders can be further complicated by the federation of an application's component across different cloud providers (Rodero-Merino, 2012). Some cloud-specific approaches (Santos, 2009) introduced a third external trusted authority to guarantee that the cloud provider could not gain access in the deployed VMs. Enterprises are often required to prove that their security procedure conforms to regulations, standards, and auditing practices regardless of the location of the system at which the data resides. Achieving auditability without sacrificing performance is yet to be accomplished. Auditors should be independent third parties, which are different from current practice in which cloud providers record and maintain their own audit logs. To change this, several efforts are under way, for example, the CloudAudit (www.cloudait.org) is developing an API which supports "audit, assertion, assessment, and assurance for cloud providers".

8. References

Archer J., Boheme A., Cullinarie D., Puhlmann N, Kurtz P., Reavis J. (2010). Top threats to cloud computing. Technical Report, *Cloud Security Alliance*. Retrieved from http://www.cloudsecurityalliance.org/topthreats.

Farber, D. (2008). Cloud computing on the horizon, CNET News. Retrieved from http://news.cnet.com/8301-13953_3-9977517-80.html.

Goldberg, R.P. (1973). Architectural Principles for Virtual Computer Systems". National Technical Information Service (NIST), U.S. Department of Commerce. Retrieved from http://www.dtic.mil/cgi-bin/GetTRDoc?AD=AD772809&Location=U2&doc=GetTRDoc.pdf.

Hotaling, M. (2003). IDS Load Balancer Security Audit: An Administrator's Perspective, *SANS Institute*. Retrieved from http://it-audit.sans.org/community/papers/ids-load-balancer-security-audit-administratorsperspective_119.

King, S. T., Chen, P. M., Wang, Y., Verbowski, C., Wang, H. & Lorch, R. (2006). SubVirt: Implementing malware with virtual machines, *Proceedings of the 2006 IEEE Symposium on Security and Privacy*, 314–327. Retrieved from http://web.eecs.umich.edu/~pmchen/papers/king06.pdf.

Kizza, J. M. (2013). *Guide to Computer Network Security*, 2nd Edition, Springer, 2013.

MacDonald, N. (2011). Yes, Hypervisors Are Vulnerable, *Gartner Blog Network*. Retrieved from http://blogs.gartner.com/neil_macdonald/2011/01/26/yes-hypervisors-are-vulnerable/.

Mell, P. & Grance, T. (2011). The NIST Definition of Cloud Computing, NIST Special Publication 800-145. Retrieved from http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf.

Metzler, J. & Taylor S. (2011). The data center network transition: Wide Area Networking Alert, *Network World*, retrieved from

http://www.networkworld.com/newsletters/frame/2011/080811wan1.html?source=nww_rss

Santos, N., Gummadi, K. P. & Rodrigues, R. (2009). Towards Trusted Cloud Computing, *Proceedings of HotCloud*, San Diego, CA, USA. Retrieved from http://www.usenix.org/event/hotcloud09/tech/full_papers/santos.pdf.

Rodero-Merino, L., Vaquero, L. M., Gil, V., Galán, F., Fontán, J., Montero, R.S. & Llorente, I. M. (2010). From infrastructure delivery to service management in clouds. *Future Generation Computer System.* 26(8), 1226-1240. DOI=10.1016/j.future.2010.02.013 http://dx.doi.org/10.1016/j.future.2010.02.013

Violino, B. (2010). Five cloud security trends experts see for 2011, *CSO: Security and Risk*. Retrieved from http://www.csoonline.com/article/647128/five-cloud-security-trends-experts-see-for-2011.

Indexing: cloud computing, hypervisor, virtualization, Infrastructure as a Service (IaaS), Platform as a Service (PaaS), Software as a service (SaaS), security

Biography:

Joseph Migga Kizza received his B.S. in Mathematics and Computer Science in 1975 from Makerere University, Kampala, Uganda, M. S. in Computer Science in 1980 from California State University, USA, MA in Mathematics from University of Toledo, Ohio, USA in 1985 nd a PhD in Computer Science in 1989 from The University of Nebraska-Lincoln, Nebraska, USA. Kizza has been with the department of Computer Science at the University of Tennessee at Chattanooga , Tennessee from 1989 were he is doing teaching and research in Social computing, Operating Systems, Computer Network Security, and Computer Forensics. Dr. Kizza has organized a number of workshops and conferences on Computer Ethics, producing proceedings and has published several books on computer ethics and network security and cyberethics. He was appointed a UNESCO expert in Information Technology in 1994.

Li Yang received her B.S. and M.S. in Finance from Jilin University, Jilin, China and Ph.D. in Computer Science from Florida International University. She is an Associate Professor at the University of Tennessee at Chattanooga. Her research interests include mobile security, intrusion detection, network and information security, access control, and engineering techniques for complex software system design. She teaches courses in Information Assurance (IA) area including computer network security, database security, biometrics and cryptography, and system vulnerability analysis and auditing. She authored papers on these areas in refereed journal, conferences and symposiums. Her work was sponsored by National Science Foundation.