Implications of Cloud Computing on Digital Forensics

Vincent Urias
801 Leroy Place
Socorro, NM 87801
1-505-284-5584
vurias@nmt.edu

John Young
P.O. Box 14174
Washington, D.C. 20044
1-202-276-4206
yjohn@nova.edu

Sherelle Hatcher
9857 Bayline Circle
Owings Mills, MD 21117
1-443-762-6356
Sherelle.hatcher@gmail.com

Abstract—**Cloud computing is a paradigm for computing services that are delivered to users over the Internet. In cloud computing, users rent rather than buy their computing resources. Cloud computing likely represents the next stage in the evolution of the Internet. But the cloud computing paradigm is still developing, with numerous unknowns and many questions open for research. One critical question that has not received much attention is security. A significant subset is digital forensics—that is, (1) the discovery of evidence remaining on a computer after a security breach or attack and (2) the use of that evidence to investigate the event and establish facts for use in legal proceedings.**

This paper discusses the impact that cloud computing will have on digital forensics. From a forensic perspective, cloud computing raises a number of concerns. Most immediate is whether or not forensic practitioners will be able to analyze the Cloud using existing techniques of digital forensics. During a traditional forensic examination, files on the storage media are examined along with the entire file system structure. But this may not be a practical model for examinations in the Cloud, where the computer is virtual, that is, where numerous heterogeneous resources, often geographically distributed, are combined. Other concerns include protecting evidence against contamination and anticipating the legal issues that will be raised by the Cloud paradigm, with its resources spread over diverse administrative and geopolitical domains. Comprehensive security services to protect not only the Cloud's resources but also the data that resides on them may need to be instituted. The open literature to date has yet to address any of these challenges.

Cloud technologies are predicted to cause a paradigm shift in digital forensic techniques. This paper discusses the application of traditional digital forensic examinations to cloud forensics.

Keywords—**Cloud Computing, Digital Forensics, Computer Forensics**

1. INTRODUCTION

This paper will differentiate between two computing paradigms—traditional computing and cloud computing. Traditional computing describes a user at a desktop workstation where data storage, applications, and computing resources reside locally. Cloud computing describes a user at a workstation or terminal connected to shared computing resources (applications, data, storage) that are accessible over the Internet. The resources can be widely distributed, and the cloud computing services can expand or contract on demand. The separate pieces that comprise cloud computing are integrated and managed for the user through an interface known as a virtual machine monitor, or hypervisor.

Cloud computing represents an expansion of today's computing environment into a virtual computing infrastructure—an expansion from networked desktop PCs to a paradigm where Cloud providers manage computational resources for multiple users who simultaneously run applications, solve computational problems, and store data. Cloud users rent rather than own the resources. Computing services are sold like a utility: Usage is metered and users pay for what they use. The Internet serves as the distribution grid and, to an extent, the computing platform.

The Cloud's primary benefits include its ability to

- create a centralized point for monitoring computer security posture and state of health
- move the management, usage, and maintenance from local hardware/software to an infrastructure that is accessed, maintained, and used from a network.

The Current State of Cloud Computing

For this paper we will define cloud computing as a shared pool of computing resources (e.g., networks, servers, storage, applications, and services) made available to users on demand through a network. These resources can be rapidly provisioned and released by the provider with minimal interaction. Because hardware is shared across multiple application instances, the network must be able to handle applications migrating from hardware to hardware, and must be configured to deal with such change without requiring human intervention.

Although cloud providers will face the challenge of delivering transparent services to the users of massive "server farms," the promise of a cost savings model that provides near limitless capabilities and accessibility to the end user is attractive for industry.

DOI: 10.5176_2010-2283_1.1.22

Already some statistical and mathematical industries as well as commercial and government entities have begun adopting public and private clouds. (Public clouds use the Internet. Private clouds emulate cloud computing, but on private networks.) The industry is beginning to understand the concepts of hosting applications and infrastructures and learning how to make the Cloud profitable [1]. As cloud computing has become increasingly mature and coherent, a few businesses have emerged as leaders in the technology. They include Microsoft, Google, and IBM. By showing their support through the promotion, encouragement, adoption, and leadership of cloud computing, the larger providers have built a foundation for the paradigm shifts of recent years. As the Cloud becomes more pervasive, the paradigm will continue to evolve. [2]

The promise of cloud computing has spurred entrepreneurial development of cloud services. A recent article lists 150 companies in its yearly roundup of cloud computing participants.[1] The services provided by these businesses are generally divided into three categories: (1) Infrastructure as a Service, (2) Platform as a Service, and (3) Software as a Service.

The subject of this paper, digital forensics, will be relevant to the infrastructure component, known by the acronym IaaS.

2. DIGITAL FORENSICS

Like cloud computing, the field of digital forensics is still in its infancy. The science of digital forensics has been described as "the process of identifying, preserving, analyzing, and presenting digital evidence in a manner that is legally accepted" [5]. Digital forensics includes both investigative and analytical techniques. Informally, digital forensics is defined as "the collection of techniques and tools used to find evidence in a computer."[2]

Digital forensics is a subset of computer security. It is the act of collecting evidence after an attack and analyzing that evidence to locate and convict an attacker. Digital forensics is complicated by the fact that successful attackers know how to cover their trails and that unsuccessful attacks often pass unnoticed.

Digital forensics is considered a science because it is a systematic, technological inspection of a computer system and its contents. Its aim is to locate and preserve electronic evidence for use in criminal investigations. Digital forensic investigation requires a level of expertise and rigorous methodology that exceeds standard data harvesting and preservation routinely performed by system support personnel.

The question is: Can we analyze the Cloud using the established tools and techniques of digital forensics? The Cloud infrastructure—with its distributed processing, storage, and resources—can be extremely complex because storage capacities can grow geometrically. From a forensic perspective this raises new questions. Before understanding the applicability of current digital forensics practices to the Cloud, we must construct a common understanding of what digital forensics entails.

Practitioners have attempted to provide some formalization to the field by defining a five-phase process: identification, acquisition, preservation, analysis, and reporting of the evidence. These phases describe a standard forensic practice that we will follow and use to organize this paper. This standard unifies many of the previous forensic protocols and provides an abstraction to the process that is not focused on a particular tool or technology, nor is it bound to a specific class of cyber crimes. Within the field of forensics are a variety of sub-domains, but they will not be addressed in this paper. Our purpose is rather to examine the larger conceptual issues that will arise in the new cloud paradigm.

Cloud Dynamics Issues

The notion of data acquisition changes when using virtual machines on the Cloud. No longer are we acquiring an operating system residing on a physical device. Consequently, it may not be feasible to physically protect against contamination of the machine through a write-blocker. In the traditional computing paradigm, the analyst physically removes the drive and takes it to a designated location to create a bit-to-bit image of the device—a copy on which to perform analyses. In the Cloud, analysts may be bound to the network where the virtual machine resides. If the network experiences a failure, it is no longer possible to conduct an investigation. Consequently, investigations will be more dependent on the surrounding infrastructure than on physical machines in the traditional paradigm.

During the experiment, a number of acquisition hypotheticals arose for which the test team did not have any definitive answers:

- What if a machine is located in location X and the examiner is in location Y and is attempting an acquisition of the device when the physical network interface card goes down?

- What will happen if the network cards fail on the server that contains the local data store with the virtual machine? If the analysts need to remove the drives, they must acquire the entire device in order to acquire the image that they were looking for, which will entail an increased amount of processing time. Additionally, analysts face legal implications with regards to other users' data that might be acquired. How are warrants handled in this situation?

- The notion of forensically sound images will be brought into question. What will the source image be in this case? Will the commercial service providers need to store the image on their systems until the case is closed, to ensure that it is indeed the genuine

[1]

[2] Caloyannides, Michael A. *Computer Forensics and Privacy*. Boston: Artech House, 2001.

> image? (If so, the examiner will be dependent on the backup strategies of the service provider.) Or will a hash of the virtual machine in question be enough to satisfy the requirement of producing source evidence in a court of law?

Despite these unknowns, however, there is a possible benefit regarding the business impact of digital forensics on the Cloud. Previously, many organizations would not bring down servers after an intrusion or event due to the financial implications. With the advent of cloud forensics, that is, live forensics, that situation may change.

Virtual Hardware Issues

The introduction of virtual hardware into the digital forensics investigation environment is a complex issue with both benefits and drawbacks. The abilities to "snapshot" a virtual machine and to preserve the contents of memory, as well as the entire state of the machine, are quite useful. However, the snapshot is a mixed blessing that opens a multitude of new and uncharted issues.

There is also the issue of virtual disks and CDs in products such as VMware's ESX, a virtual machine monitor or hypervisor. These devices can operate just like the physical addition of a hard drive to a machine. Users can specify the size and type of drive (such as IDE or SCSI), among other features. From the operating system, it would appear as though a new physical drive appeared on the device. However, if one were to snapshot and add a drive, then revert to the snapshot without the drive, there would be no evidence from the operating system that a secondary storage device ever existed. This is a sophisticated method for ensuring data are obfuscated on the Cloud. The notion of file carving, the location of deleted virtual partitions on the physical disk, and the amount of time they would reside there before the space would be reclaimed by a new virtual machine, are all new questions that will remain unanswered until more research is conducted.

Additionally, the ephemeral nature of cloud computing raises many issues regarding the lifetime of a particular device because the lifetime is no longer years or months; rather it is weeks at best. Storage needs grow as the Cloud grows. The storage issues are one of the greatest challenges of cloud computing. As demand for resources increases, the cloud provider's ability to store all of a particular users' information for weeks, or even months, becomes economically unfeasible. As space is reclaimed, forensic evidence is lost.

3. EXPERIMENTAL SCOPE

This section will focus on the potential to leverage some of the Cloud's capabilities to create an advantage for forensic analysts.

The cloud computing environment includes three primary elements: platforms, software, and infrastructure. All three are configured and delivered as services. In this paper, and in our experiments, we limited our work to that part of the cloud relevant to digital forensics—the infrastructure. But note that the cloud infrastructure is virtual, not tangible. Cloud infrastructure is organized to ensure scalability and make efficient use of resources, not to facilitate forensic investigations. In the Cloud, infrastructure is configured with numerous separate components that can reside anywhere on the cloud provider's network. Consequently, forensic evidence might be widely dispersed over many devices and domains.

The cloud paradigm benefits users but complicates matters for forensic investigators. Cloud technologies will force a shift from traditional forensic techniques.

In the near term, the forensic process in the Cloud is likely to rely on the examiner's knowledge of the technical aspects of the specimen and an understanding of the case and the law.

In the longer term, the success of a forensic examination will be strongly dependent on how the Cloud is finally deployed.

Fortunately the Cloud infrastructures have certain common attributes.

- Computing infrastructure are standardized and scalable
- User focus on the usage of the Cloud
- Cloud provider takes care of keeping it running
- Self-service based usage model
- Users manage their own application
- Minimal or self-managed platform
- Infrastructure doesn't require a lot of care and feeding to keep running.

These attributes provided a starting point for developing a set of cloud forensics experiments. To conduct those experiments, we needed to build a cloud environment and select cloud management software to operate it.

Selection of Cloud Management Product

An analysis of the cloud computing market revealed that the VMware vSphere suite of products (described in the following section) has been adopted by numerous Fortune 500 companies as well as government customers. These high acceptance rates suggest that industry and government will be using vSphere (and other VMware products) in numerous venues.[3] Thus, analyzing this product will potently yield far-reaching effects. As a result, the researchers in this study chose

[3] Bruzzese, J. Peter. "The Hypervisor Wars," *Infoworld*, 9 September 2009. According to IDC's Worldwide Quarterly Server Virtualization Tracker, the majority of IT shops using virtualization are working with VMware products. (http://www.idc.com/getdoc.jsp?sessionId=&containerId=219011)

vSphere to be the model by which to analyze all the problems. However, all the problems and conclusions which are drawn should ideally be tools/technology agnostic.

The following reasons formed the basis for the choice:

- Ease of acquisition
- VMware offers free 30-day trials on all the software components needed to build a cloud computer
- VMware is well-known and has been one of the market leaders in virtualization; therefore, many resources are available for this product.
- Ease of use: Many blogs and much documentation exist to aid in resolving problems with building and maintaining the vSphere cloud computer.

Product Description

vSphere is a suite of products that enables a virtual infrastructure. The vSphere virtualization stack was released on 21 May 2009 by VMware of Palo Alto, CA. The manufacturer promotes vSphere as a datacenter virtualization platform. vSphere manages collections of infrastructure (e.g., CPUs, storage, and networking) as a pool of resources to call on in order to complete jobs—similar to the performance of commercial operating systems that take the pool of local resources (CPU, memory, storage, networking) and manage them as a single entity. vSphere falls under the IaaS paradigm.

vSphere Components

The vSphere hypervisor, known as ESX, allows multiple operating systems to share a single hardware host. This virtualization layer runs on physical servers that abstract the various resources (processor, memory, storage) into multiple virtual machines. Two versions of ESX are available: ESX and ESXi. The former is an enterprise-level product. The latter is a streamlined, free version of the product.

Important features of the vSphere stack include the following:

- VMware ESX 4.0 contains a built-in service console. It is available as an installable CD-ROM boot image.
- VMware ESXi 4.0 does not contain a service console. It is available in two forms: (1) VMware ESXi 4.0 Embedded and (2) VMware ESXi 4.0 Installable. ESXi 4.0 Embedded is firmware that is built into a server's physical hardware. ESXi 4.0 Installable is software that is available as a CD-ROM boot image. Users install the ESXi 4.0 Installable software onto a server's hard drive.
- VMware vCenter Server is the central point for configuring, provisioning, and managing virtualized IT environments.

VMware vSphere Client is an interface that allows users to connect remotely to vCenter Server or ESX/ESXi from any Windows PC.

4. EXPERIMENTS

To test the vSphere product, we devised a series of experiments. Our experiments were conducted across several different machines with differing architectures in order to evaluate vSphere on a variety of platforms. The experiments during this study were conducted entirely on a private, stand-alone cloud.

Additionally, we hoped the ability to run vSphere on heterogeneous pieces of equipment with different memories, computational powers, and disk space would demonstrate the potential for heterogeneity in the cloud environment.

We also used various versions of the VMware software to determine scalability and backward-compatibility within the ESX/ESXi product line.

Experiment Topology

Several difficulties were encountered in setting up the environment, particularly related to the hardware specifications for vSphere. Previously, all VMware products (ESX/ESXI 3.0, 3.5 u1 – u4), even the most recent release 3.5 u4 issued in March 2009, had been installable on 32-bit hardware. In prior versions of ESX, 32-bit hardware was supported in "legacy" mode. However, after the release of vSphere, support for 32-bit hardware dropped and only 64-bit hardware was supported.

Additionally, there was a need to have the Intel-VT chipset for vSphere to install properly on the various machines. Numerous different techniques were attempted to enable ESX/i 4.0 support for 32-bit hardware, ranging from changing parts of the installer to upgrading from an ESX3.5 machine to ESX4.0, with no success.

After supported hardware was used, the installation and configuration of both ESX and ESXi was quite straightforward. Because ESXi (the free version of ESX) has a reduced functionality, the installation walked through everything and worked quite well. However, because of the heterogeneity of our environment, some of the core features (such as HA) didn't work well in non-uniform environments. As such, many of the add-on packages were not utilized.

Some seemingly random issues were encountered while conducting the experiments:

- The clients would lock-up or become unresponsive after some use. The problems were not apparently deterministic (in terms of the activities or the applications that were running).
- Users could not access the virtual machines or we were unable to power ON/OFF or edit the virtual machine settings.

These problems were mostly attributed to the immaturity of vSphere. However, they are issues which may be encountered or may be leveraged as an operating artifact for other activities.

The test topology was as follows:

- A combination of ESX 4.0, ESXi 4.0, and ESX 3.5 Update 4 infrastructures
- vCenter was used to manage the clouds resources and to instantiate virtual machines
- An NFS share was used as a datastore (hosted by an Ubuntu 9.04 workstation)
- A combination of distributed switches and virtual switches were used
- A combination of local datastore and NFS datastore was used
- A variety of 64-bit and 32-bit hosts were used
- A variety of hosts (Windows XP, Windows Vista, and Windows 7 with differing service packs, and Ubuntu) were used

During the course of the experiment, the number of hosts and the location of the virtual machines changed; however, the basic topology remained consistent.

5. POTENTIAL SOLUTIONS

This section presents potential solutions to some of the problems raised in the experiments. We will discuss some areas where development can be focused to address the anticipated problem. These areas include logging, hypervisor forensics, network forensics, and digital forensics on demand (Forensics as a Service).

Hypervisor Forensics

The notion of the virtual machine sitting on top of a lightweight hypervisor is a relatively new paradigm that forensic practitioners are beginning to address. Traditional forensic techniques, based on assumptions that the file-system was directly interacting with the hardware through an abstraction, afforded the forensic practitioner the ability to assume that there was nothing controlling the application below the file-system. This is not the case when using virtualized technologies such as Zen or VMware products. These hypervisors have the ability to covertly add, remove, and/or modify hardware and software on the virtual machines on the fly, thus introducing questions concerning the validity of data being analyzed on the virtual machine. Activities like adding or removing virtual hardware would be recognized by the hypervisor but not the virtual machine per se, thus necessitating development of a hypervisor-level forensic tool (which sits at the hypervisor level) for logging, verifying, and storing these changes to the virtual machines and archiving the information for use during an investigation.

Network Forensics

Network forensics is a subset of digital forensics that focuses on collection of network data that may be useful as evidence (files that have been transferred, among other activities) and the analysis of packets to reconstruct events. Network data will play a significant part in cloud computing forensics. With the absence of local media to transfer files, the network will be one of the only methods to transfer data among machines and as such will leave forensic network evidence of such transfers. Although the idea sounds simple, the collection, preservation, and analysis of network data pose a complicated problem. The amount of data generated on the network can be more than gigabytes per day. On the Cloud, data generation will most certainly be terabytes if not hundreds of terabytes per day. The storage capacity necessary to archive this amount of data (perhaps for years) is inconceivable.

Another challenge is the ability to trust the network data being received, since the Internet and routing protocols are inherently anonymous and much information from the packets can be spoofed—such as the MAC address, source IP, and destination IP. All are mutable fields that are simple to alter. Users can also encrypt traffic using methods such as SSH-tunneling.

Nevertheless, the Cloud provides an entirely new venue and a potential wealth of forensic artifacts, although formidable challenges face investigators before those artifacts are proven valid in the legal arena.

The network forensics principles currently in use are the ability to address evidence collection, information storage, and retrieval in combination with the attack attribution abilities.

Several researchers have published papers on the topic of implementing network-based forensic techniques [10; 11; 12; 13; 14]. However, none of these techniques have been implemented in production environments. The researchers have identified the inherent ability to use, manipulate, and analyze network information that is collected in meaningful ways; however, they have not provided methods for conducting analysis in real time or on large datasets.

Real-Time Analysis

There is an urgent need to develop smarter, more powerful, distributed tools to analyze Cloud data in a real-time sense. Although strides have been made on tackling the network forensic problem, with researchers working in parallel, little progress has been made. New strategies and computing techniques must be devised to address this issue, maybe even using the Cloud itself to analyze the data. Existing tools such as Map-Reduce or Hadoop may provide the flexibility, modular framework, and computational resources necessary to tackle these problems. However, someone must first test and vet these tools.

If a company such as Amazon were to tap or log all the traffic coming in and out of its network, an immense amount of information could be discerned. (Let us ignore the issues with storage and the vast amount of information that will be passing through the network.) First, there is an increased ability to assign attribution. By looking at the entire set of network information, it becomes possible cross-correlate data related to who is communicating with whom and what virtual machines and files a particular user is transferring to and from the Cloud. In addition, there is an increased ability to use the file system information between the EC2 time stamping and the hashing along with the net flow data to create a stronger case for the occurrence of a particular event. Finally, if the data have been removed from the virtual machine, thus removing any file-system forensics, it is completely reasonable to use network dumps to try to reconstruct events.

Due to the ephemeral nature of the data stored on the virtual machines, a claim can be made that if the tools were to be vetted when moving to the Cloud there will be a stronger reliance on network forensic tools. These tools will provide not only a historical, imputable record of events but also the ability to cross-correlate and begin to address how to gather evidence in the cloud.

6. CONCLUSION

During the course of the research described in this paper, normal disk forensics techniques were explored that would no longer work in the expected manner. Several solutions were proposed to this problem. However, the purpose of this paper was to bring to light the issues of forensics that would need remediation.

It is clear that traditional digital forensics will be insufficient as the Cloud grows. Before cloud computing becomes more widely accepted, it is essential that we identify the emerging challenges and begin to develop solutions or mitigation techniques. Such efforts are essential if we hope to preserve the basic tenants of computer security.

Piecing together forensic data after the fact from distributed sources not intended to be used for that purpose would be a frustrating and unsatisfactory process. Tools must be developed to ensure that the necessary information can be securely collected and remain available if an attack is successful.

ACKNOWLEDGMENT

Our thanks to ACM SIGCHI for allowing us to modify templates they had developed.

REFERENCES

[1] Bowman, M., Debray, S. K., and Peterson, L. L. 1993. Reasoning about naming systems. ACM Trans. Program. Lang. Syst. 15, 5 (Nov. 1993), 795-825. DOI= http://doi.acm.org/10.1145/161468.161471.

[2] Ding, W. and Marchionini, G. 1997 A Study on Video Browsing Strategies. Technical Report. University of Maryland at College Park.

[3] Fröhlich, B. and Plate, J. 2000. The cubic mouse: a new device for three-dimensional input. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (The Hague, The Netherlands, April 01 - 06, 2000). CHI '00. ACM Press, New York, NY, 526-531. DOI= http://doi.acm.org/10.1145/332040.332491

[4] Tavel, P. 2007 Modeling and Simulation Design. AK Peters Ltd.

[5] Sannella, M. J. 1994 Constraint Satisfaction and Debugging for Interactive User Interfaces. Doctoral Thesis. UMI Order Number: UMI Order No. GAX95-09398., University of Washington.

[6] Forman, G. 2003. An extensive empirical study of feature selection metrics for text classification. J. Mach. Learn. Res. 3 (Mar. 2003), 1289-1305.

[7] Brown, L. D., Hua, H., and Gao, C. 2003. A widget framework for augmented interaction in SCAPE. In Proceedings of the 16th Annual ACM Symposium on User interface Software and Technology (Vancouver, Canada, November 02 - 05, 2003). UIST '03. ACM Press, New York, NY, 1-10. DOI= http://doi.acm.org/10.1145/964696.964697

[8] Spector, A. Z. 1989. Achieving application requirements. In Distributed Systems, S. Mullender, Ed. Acm Press Frontier Series. ACM Press, New York, NY, 19-33. DOI= http://doi.acm.org/10.1145/90417.90738

Vincent Urias is a student in the School of Information Technology at New Mexico Institute of Mining and Technology. He is doing research in the field of network security.

John W. Young Jr. is an adjunct professor in the School of Information Technology at Marymount University. He is currently working on his PhD in Information System at Nova Southeastern University. He is doing research in cloud computing security.

Sherelle Hatcher is a graduate student at Stevenson University studying in the School of Forensics Information Technology. She is doing research in the field of digital forensics, network security and network vulnerability.