Overview of the Forensic Investigation of Cloud Services

Jason Farina, Mark Scanlon, Nhien-An Le-Khac, M-Tahar Kechadi
UCD School of Computer Science and Informatics,
University College Dublin, Belfield, Dublin 4, Ireland
Email: jason.farina@ucdconnect.ie, {mark.scanlon, an.lekhac, tahar.kechadi}@ucd.ie

***Abstract*—Cloud Computing is a commonly used, yet ambiguous term, which can be used to refer to a multitude of differing dynamically allocated services. From a law enforcement and forensic investigation perspective, cloud computing can be thought of as a double edged sword. While on one hand, the gathering of digital evidence from cloud sources can bring with it complicated technical and cross-jurisdictional legal challenges. On the other, the employment of cloud storage and processing capabilities can expedite the forensics process and focus the investigation onto pertinent data earlier in an investigation. This paper examines the state-of-the-art in cloud-focused, digital forensic practises for the collection and analysis of evidence and an overview of the potential use of cloud technologies to provide Digital Forensics as a Service.**

I. Introduction

Cloud Computing has become a household term in recent years referring to numerous different commercial and consumer oriented products and services. Cloud computing can be used to provide high availability to the servers of an organisation, off site replicated backup solutions, tools for use in disaster recovery and business continuity as well as instantly scalable resources that only require capital expenditure for as long as the resource is required. Once the processing power, storage or backup is no longer needed the additional resources allocated can be handed back and, depending on the negotiated contract, no longer be considered a business expense.

Cloud computing is not just for businesses and large organisations capable of purchasing and hosting their own server farms and distributed resource pools. The availability of cloud based services to the consumer has augmented the processing power and storage of mobile devices. Many client-server based applications are in reality cloud-based Platform-as-a-Service models transparently presented as a single entity to the application. Online storage utilities are thought of as a single block of storage used by the end user to upload and download files where in actuality they are virtualised, distributed containers allocated to end users based on the service level they have signed up for.

This incredible complexity underpinning the mechanism to provide dynamic resource allocation and management is invisible to the end user. The same, however, cannot be said for the digital forensic investigator. In some cases an investigator will be aware of the requirement of cloud-based forensic practises should they be aware of the scope of the investigation prior to the undertaking. In others it is only in the course of investigation that they discover that the 32GB storage on a suspect tablet is actually a front-end for a multi-gigabyte or terabyte container provisioned on a cloud. The tablet could just be used as an entry point for an entire network of servers and clients hosted across multiple datacentres distributed across the world. A warrant for a simple search and seizure of a tablet could turn into an international jurisdictional nightmare. One in which there is no physical access to any of the systems involved and no way to ensure the co-operation of the only entity that knows the inner workings well enough to produce forensically sound data for analysis.

Within the EU, a centralised approach to the investigation of European cybercrime cases can greatly ease the administrative overhead for national law enforcement agencies, such as Europol's European Cybercrime Centre (EC3) [1].

A. Aim and Contribution of this Work

In this work we attempt to provide:

- An account of the current technical obstacles and legal impediments to the comprehensive evidence locating and gathering process from cloud computing sources.
- An identification of areas of research either in progress or yet to be undertaken that could be of benefit to forensic investigations involving cloud-based evidence.
- A connectivity between security practices for the cloud and how these could feed into Forensic processes.
- Identification of areas where the cloud can help enable forensic investigation rather than just be seen as a challenge to be overcome

II. Background Reading

A. Cloud Computing

There have been numerous attempts at defining with precision what the term "Cloud Computing" engenders. The most reliable and widely accepted definition of cloud Computing is provided by the National Institute of Standards and Technology (NIST) [2], which states that *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.* This definition describes the way in which a cloud computing system operates rather than defining an exact technology,

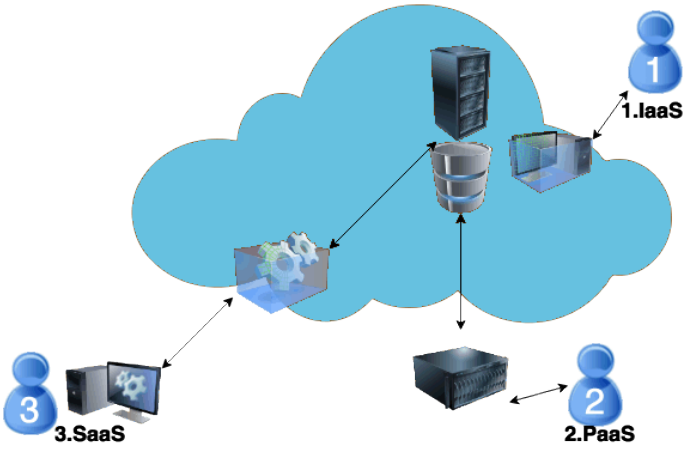

Fig. 1. Cloud Service Architectures

architecture or specific set of services provisioned by the vendor. This generalisation is indicative of the challenges facing the forensic investigator undertaking any investigation that involves a system that is cloud provisioned or that utilises a cloud provisioned service.

NIST divides cloud computing into three complimentary service models, as depicted in Figure 1 and outlined in the following sections [2]:

1) Infrastructure as a Service (IaaS): IaaS, represented as `User #1` in Figure 1 refers to the providing of virtual hardware, such as servers, storage and networking which allows customers to build virtual infrastructure which mimics the traditional physical computer hardware. The most popular IaaS provider is Amazon Web Services (AWS). Their IaaS offering includes their Elastic Compute Cloud (EC2) on-demand cloud computing instance offering and their Storage as a Service (StaaS) offering, Simple Storage Service (S3).

2) Platform as a Service (PaaS): PaaS, `User #2` in Figure 1 operates at a layer above the visualised raw computing hardware. PaaS provides methods for tools to be developed that easily interact with services such as databases, web servers and file storage. These services are abstracted from the underlying physical storage space constraints, replication and redundancy planning and load balancing. Some examples of PaaS services include Google App Engine and Force.com. In 2014, Almulla et al. found that no research had been conducted on PaaS forensic investigations [3].

3) Software as a Service (SaaS): Seen as the most likely entry level of cloud-based computing for most users and businesses, SaaS is defined by NIST as *"the capability provided to the consumer is to use the provider's applications running on a cloud infrastructure"*. Perhaps the best known SaaS provider is Salesforce.com which specialises in sales and marketing suites hosted in their data-centres but accessed via licensed connections from the customer's systems. Clients leverage the capabilities of Oracle and SAP back end services without having to pay for the installation, licensing or maintenance for these software and hardware solutions. Most recently, Salesforce.com have added Data Analytics as a Service to their product line. Most SaaS models are advertised under a "pay-for-what-you-use". This option is represented as `User #3` in Figure 1.

A second type of SaaS is also present in the form of Application Service Providers (ASP) which sub-licenses software to customers who can access it on the ASP's hardware. An example of this can be seen in the services provided by some web hosting providers where email servers running Microsoft Exchange can be rented with pricing set by the number of licenses required. The ASP provides the access portal to the server and the storage space but leaves the application administration to the client. The client however has limited or no control over the infrastructure or underlying server components and may require support from the ASP to make any degree of significant changes beyond content management.

Mobile cloud computing affords mobile users with the additional processing power, storage and functionality not normally available locally on the device itself [4]. With the advent of always-on data connections in the mobile computing world, many forensic investigations of mobile devices can be aided with cloud digital evidence gathering. Cloud recovered digital evidence of mobile backups can contain text messages, photos, videos, application data, etc. The recovery of this mobile backup data may be the only method of accessing this data from a compromised mobile device due to encryption or device degradation.

The practice of cloud computing forensics requires a blend of many different digital forensic skills depending on the type of cloud under investigation. In order to provide flexibility and scalability, cloud systems are usually built on a virtualised environment [5] with dynamically allocated resources. To efficiently utilise the underlying hardware, virtual environments, even those not used for cloud systems, are usually over-allocated. When an investigation involves an IaaS, StaaS, PaaS or SaaS the methodology for investigating a network, file-share, workstation or an application becomes more difficult to apply as the scope of the investigation becomes a series of potential maximums as opposed to an actual value.

In a cloud system the amount of hard disk space allocated to a volume will expand and shrink dynamically, the amount of RAM available in the resource pool will vary and `read for use` + `in use` will not always tally to `total RAM available`. Cloud systems are usually designed around the virtualisation practise of `over subscription`. This is the practise of taking the resources designed to be distributed between 10 systems and instead dividing them between more (maybe 15 or 20) in order to take advantage of the fact that any single system will not use all of its resources at any one time. If each of 10 systems only uses 50% of the storage capacity available to it, there would be enough storage capacity for another 5 similarly sized systems. In order to allow for spikes in activity and outliers in usage, the Cloud Service Provider (CSP) can opt to sell 4 additional systems and leave enough storage space slack to compensate. The same tenet can be

applied to RAM and network bandwidth. Generally speaking, it cannot be applied to virtual CPUs unless the CSP offers the option of multiple CPU availability – in which case, unused clock cycles on an idle CPU can be assigned to another VM that requires it.

III. Overview of Cloud Security

While security generally resides at the opposite end of the investigative spectrum to forensics, security practices, if implemented correctly, can provide evidence that can support the forensic procedure in recreating events and in some cases proving capability of access. The ability to implement security effectively depends on an organisation's level of control over the underlying hardware of the cloud system they utilise.

NIST also describe three classifications of architecture for the deployment of cloud services which must be taken into consideration in addition to the levels of service. These three architectures, by their nature, introduce levels of security concerns.

- `Public Cloud` - the Public Cloud is a cloud service that is hosted completely external to a client's infrastructure. All storage, infrastructure, platforms and data are housed on systems owned entirely by a third party provider. This introduces three issues of security that will need to be addressed by the Client. First, the public cloud system exists outside of any security measures or policies enabled in the clients own premises. Second, the data must be transferred between the client site and the hosting company, usually across a connection that traverses untrusted pathways and finally the service is hosted on systems that require security provided by a third party and as such must also be relied upon to provide security for any data at rest or system access and hardware configuration.
- `Private Cloud` - the private cloud is a cloud system hosted and operated entirely within the boundaries of the client's site or systems. While the private cloud provides security in the form of trusted operators and systems that fall under the policy scope of the client it also introduces insecurity in that it lacks the high availability provided by the scale of dedicated CSPs.
- `Hybrid Cloud` The compromise between these two ends of the architecture spectrum is the hybrid cloud that involves select data, systems, or services run from the cloud that interact with systems, services and data hosted within the client's domain. In effect the organisation extends its DMZ to include systems hosted at a third party facility, not unlike the way co-located hosting of web services is already managed. This hybrid model increases the security of the cloud portion of the architecture but increases the attack surface of the organisation and brings a shared resource within its security perimeter.
- `Community Cloud` NIST also define a deployment strategy as a "Community Cloud" where a group of like-minded individuals or organisations with similar goals and requirements pool resources to host a form of semi-public/semi-private cloud infrastructure. Though more controlled than a full public cloud there still exists a requirement of trust when dealing with co-cloud clients and those tasked with maintaining the system. for Forensic purposes this type of cloud would be considered either Hybrid or Public depending on how large the participant pool was and how tightly controlled and granular the access.

In all of these scenarios the level of security is a function of the level of trust that can be placed in the third party CSP and how much of the organisation has integrated the cloud structure into their own system architecture. When allowing any external service traffic in or out of their own security perimeter the organisation must first consider how much faith can be put in the CSP's ability to correctly manage, maintain and implement segregation of data and access in a dynamic and often overlapping environment. Usually the level of service provided is agreed as part of the initial contract between the parties and is monitored through the use of a Service Level Agreement (SLA). SLAs however, generally do not deal with security or forensic concerns and instead concentrate on the service provision in the form of availability (the popular 5 9s uptime) and usability (processing times based of data volume to be processed). Any security and log monitoring is usually deemed the responsibility of the client [6]. This onus of reporting poses an issue as, unless deliberately implemented as part of the SLA or negotiated as part of the service provided, most cloud solutions that provide PaaS or SaaS and in some cases StaaS do not provide user access to the log files that record user access to the data stored on the cloud. In [7], Popa presents an addition to cloud services that provides this audit capability to the end user. This application, Cloudproof, manages an Access Control List (ACL) for each block of data residing on the remote systems. At the end of a custom interval, termed an epoch, the end user performs an audit of the ACL activity.

Using CloudProof, access can be granted in the form of Read-Only or Read-Write. Read-Only access consists of the key to decrypt a secure stream cypher that protects the data at rest and in transit. Write enabled access utilises a form of public/private key pair system (denoted as `verification` / `signing` key pair in [7]) allows a user to alter the data stored with each action verified by the ACL and each write similarly checked against the `signing` key. The majority of workload involved in this process is offloaded to the cloud system itself and results in a reported overhead of approximately 17% depending on the level of logging and checking enabled. This audit log can then be utilised in a forensic examination of the system to determine capability (access level) and establish a timeline based on the recorded access times.

Systems like CloudProof, while aimed squarely at security issues, can be useful to aid forensics in performing post event reconstruction. They are also useful for ensuring provenance of evidence. In [8] Lu et al. present a system of digital signing performed by a user whenever they write or change

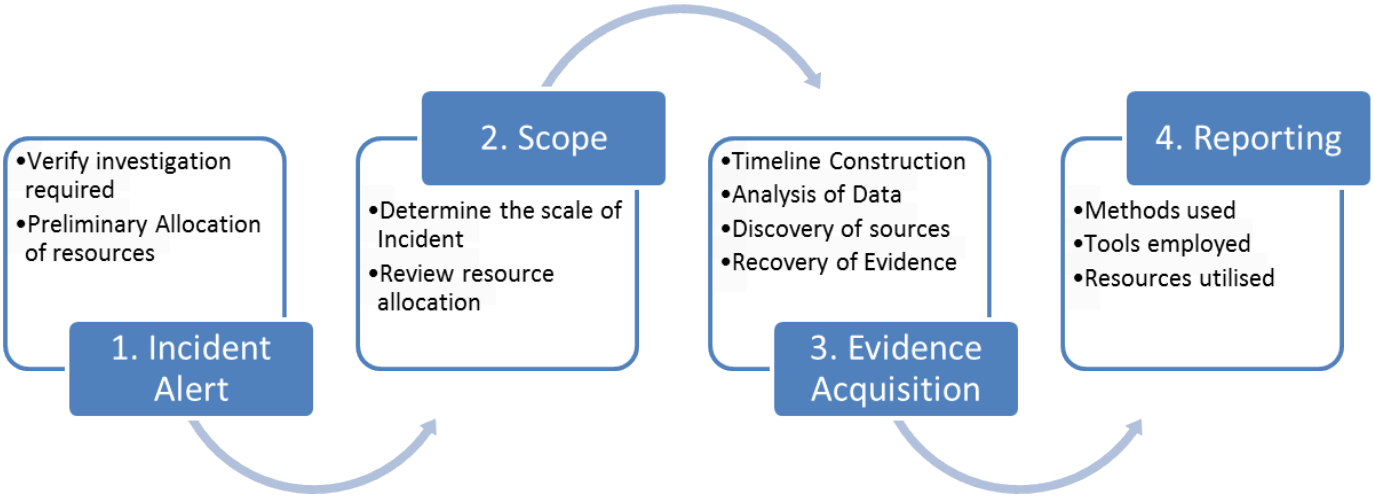

Fig. 2. A generalised forensic process. Stage 3 is repeated as required until the investigator is satisfied or no more evidence can be discovered or recovered

any data on a cloud system. Through a series of security based game designs, Lu shows that it is possible to utilise a Diffie-Hellman Key exchange variant (The Decisional Bilinear Diffie-Hellman) to provide logged and verifiable access to a cloud system while maintaining a level of anonymity. Lu goes on to prove the levels of security provided through a series of simulated attack vectors.

Another approach to security using logging was presented by Zawoad et al. [9]. This secure logging service was suggested as an optional SLA offering to log all guest interactions and store them external to the guest system itself in an encrypted format. The suggested data collection is based on Snort (a popular open source network based intrusion detection system (NIDS)) rules which would monitor and parse traffic and log only those events that triggered a warning or alert. Logs belonging to the systems of each client are collected by an aggregator known as the log accumulator and sent to accumulator storage. Each agent in the process (CSP, client and auditor) is used as a balance to the other two and Proof values can be produced by any party to support or invalidate a claim. The use of the VMM as the logging agent and providing the logs to the relevant parties through a non-guest related API ensures that logging cannot be disabled through a compromise of the guest OS.

IV. Considerations for Forensics in Cloud Environment

Cloud stored evidence can often be split across multiple different devices, frequently resulting in the evidence being impossible to find [10]. In [11], Grobauer and Schreck perform a short gap analysis of incident response procedures when applied to a cloud environment. They evaluate the current accepted best practise approach in stages and identify the primary issues that a responder will face. The same process can be applied to the standard digital forensic methodology.

The accepted methodology of a forensic investigation, as depicted in Figure 2, can be described as four basic steps as outlined below:

1) Verification - The forensic investigator must first verify that an incident has taken place.
2) System Description - The investigator must determine what systems must be included within the scope of the investigation.
3) Evidence Acquisition - The evidence must be recovered and preserved in a forensically sound manner. Evidence should be collected in order of volatility to ensure as little data as possible is lost due to the passage of time. This is further broken down into four steps, repeated until an exhaustion of evidence sources.
    a) Timeline Reconstruction - Using the evidence available the investigator must reconstruct the sequence of events that lead up to and immediately followed the even that triggered the investigation.
    b) Data Analysis - Recovered relevant data is analysed to provide details that will aid in timeline reconstruction and identification of the transgressor.
    c) Discovery - Based on the current information available the acquired evidence is inspected to determine if further information can be extracted.
    d) Recovery - Data that has been destroyed or otherwise obfuscated must be retrieved so that it can be inspected.
4) Reporting - The findings must be reported in a manner that allows for verification and should include a description of the tools used at all stages of the investigation.

At each of the stages in this generalised forensic methodology the cloud environment presents challenges that can be technical, physical or legal in nature, or a combination of the three.

Popa's [7] treatment of enabling access security through logging and ACL stems from a built in mistrust of the CSP

born of diverging interests. Should a data breach occur, the client will most likely try to find fault with the system in some way while the provider will try to avoid negative publicity and damage to their reputation by looking for fault with the client access controls and key management. In traditional forensics an investigator is presented with a suspect system and, depending on the perceived complexity of the task, proceeds with the best practise method of investigation (either live forensics in a case where volatile evidence may be of importance or post mortem forensics when data can be analysed at a later time). Usually, again dependant on the individual case, the instructions to the investigator will include a scope defined by an individual and/or by a location set out in a court issued warrant of some description. For example, a warrant issued to investigate all electronic storage devices operated or accessible to suspect A at his place of work would include a desktop workstation, possibly a laptop or portable storage devices, email archives and network file shares but would not include the systems of co-workers or secure file servers to which the suspect does not have access. This scenario, when applied to the cloud environment presents several challenges both legal and in practicality.

A. Legal Responsibility

From a legal perspective, any given cloud is usually owned by a single entity that is not usually the suspect. Data stored on a cloud system is usually constantly in motion across multiple data centres around the globe. An exception to this is data tethering - a practise where an organisation can stipulate as part of the SLA that data does not replicate to data centres located outside a geographic or geopolitical area, this can be a requirement for Public bodies storing low level but still government owned data. The reasoning behind this practise of data dispersal is twofold. First it ensures that not all data is stored in any one place which increases data redundancy in case of disaster and the multiple locations allow for simultaneous access which decreases sequential read times for large data sets. Secondly, it facilitates the ability for cloud providers to "chase the moon" where data stores are deemed active when they are in an area that has entered a cheaper power rate time period such as overnight energy rates in many European countries. This dispersal raises the issue of location. The data can be in many data stores at any one time, sometimes the same data is in multiple stores or is in transit. A court order issued in the jurisdiction that contains one data store may not be applicable to the jurisdiction that contains another. It also would not be feasible for an investigator to go to each data store to perform data collection.

This data collection can also give rise to legal issues. In the majority of cloud services, any data is stored in multi-tenancy virtual hosts which share data stores attached to each host via networks or fabrics depending on the architecture. Any attempt to physically connect to a data store or virtual host system to retrieve data directly will run a risk of modifying data that is outside the scope of the investigation insofar as belonging to a system that is not owned or operated by the suspect named in the warrant. This can also lead to breach of privacy should any incidental data be copied along with whatever part of the suspect data can be identified. For this reason, the investigator requires the assistance of the CSP to use the systems designed to collate and present the client data to produce a forensically sound copy of the data stored in the cloud systems.

This gives rise to legal questions concerning who should or must be involved in an investigation and what are the respective responsibilities of the key parties defined by NIST as [12]:

- The Cloud User - The individual or organisation that subscribe to the cloud resources under investigation.
- The Cloud Service Provider - The entity or entities responsible for the design, supply and maintenance of the cloud resources
- The Cloud Broker/Agent - The legal entity with which the cloud User has negotiated a contract for the procurement of cloud resources as well as the manager for the use, performance and delivery of cloud systems to the User.
- The Cloud Auditor - A party that can conduct independent assessment of the cloud services and systems. An auditor should assess in terms of SLA adherence, security controls, privacy impact and performance.
- The Carrier - The ISP of the cloud user and/or the CSP.

In his paper, Chen [13] identifies the need for legal redefinition when dealing with cloud systems in order to apply legal roles and responsibilities to each of the key parties and to properly define legal evidence standards when recovering evidence from a cloud environment. In addition Chen refers to the requirement for a clarification on the expectation of privacy laws for data that is hosted with a third party. This would only apply to the public, hybrid and community cloud deployments however as the private deployment is still within the perimeter of the organisation. Chen also notes that there needs to be a clarification in the definition and potential ramifications for cloud-based crimes such as VM hopping (also known as VM Hyper-jumping) or accessing cloud user owned data without consent or knowledge (or legal grounds).

NIST have compiled a list of 65 concerns involving forensic investigations that require some degree of interaction with data stored in a cloud [14]. Of the concerns recorded and currently awaiting public opinion and feedback before finalisation, 14 are classed as legal in nature with 3 being directly related to jurisdictional issues that arise from the distributed nature of data storage in cloud infrastructures. Other concerns raised are the reliance on the internal systems and resources of the CSP to provide timely data recovery and the lack of legal requirement for metadata or log file retention.

B. Remote Cloud Forensics

Remote forensics is the retrieval of data hosted by a third party on a system that may or may not be directly controlled by the owner of the data. Forensics performed on remote systems is usually performed only when the investigator cannot get physical access to the system hosting the data. This may be as a result of not knowing where the system is (as can be the case

in investigations involving anonymising networks), or for logistical reasons such as the CSP's practice of data distribution. The remote gathering of digital evidence can be beneficial in extending the digital evidence acquisition window, where local data is no longer recoverable due to corruption, over-writing or encryption [15]. Such remote acquisition requires the same attention to detail as regular cloud forensic investigations as care must be taken to ensure that only the suspect data is extracted from the remote location. Anything more may be considered an unwarranted search or even a breach of an individual's privacy. As with cloud forensics, many researchers recommend utilisation of a recognised client to perform the authentication and retrieval while others argue that this brings an inherent additional concern of the accuracy and forensic soundness of a third party API or application.

Similar question and concerns can be applied to the need for clearer legislation and better defined legal requirements with respect to remote forensics in the case of "seedboxes", remotely hosted servers almost exclusively for the purpose of facilitating peer-to-peer (P2P) file transfers, most often involving BitTorrent but other P2P protocols have been used also. These seedboxes are seen as a way around using a home IP address for downloading and some providers[1] advertise the fact that they securely delete all of their traffic logs on a daily basis as a form of protection for the activities of its customers. It is important to note that not all seedboxes are cloud-based services but those that are set a precedent for enabling anonymity in the cloud for potentially illegal activities.

C. Third-Party Compliance Issues

Cloud systems, with the possible exception of `Private Cloud`, are not centrally located. Data and resources are spread among LUNs across a series of one or more internetworked data stores made up of one or more storage arrays or data silos. This architecture ensures that any forensic investigation requires access to software capable of reading the environment and of successfully extracting data from it. For this an investigator requires the use of the CSP software or of a software suite proven to be accurately capable of interacting with the CSP systems. To operate the software, the investigator must know how the CSP systems operate to a degree of knowledge that would allow him or her to know when incorrect data has been returned as the result of an eDiscovery or retrieval operation. The safest alternative is to request that the CSP provide the data from their systems themselves in a format that can be utilised by the investigator. This is currently the way forensic investigations involving ISPs are carried out. A request for information is made under an agreed framework and the required data is returned, if held, by the ISP. For example, in an investigation where an IP address must be tied to an account holder, in the UK a RIPA request (as part of the RIPA act 2000) is performed where an ISP is required to provide the name of the account an IP address was allocated to at a specific time stamp.

The legal issues and uncertainties raise concerns for forensic investigations involving cloud systems. Civil investigations may not have the capability to perform an investigation without the co-operation of the CSP itself which may not be forthcoming without legal persuasion or if the SLA with the Client does not include any allocation of responsibility in the event of a security issue or breach.

Even with the full co-operation of the CSP there are legal considerations to be taken into account that may have a bearing on the quality of evidence that can be presented and its admissibility in legal proceeding whether civil or criminal.

Is the technician providing the data acting as an agent of the legal system in this matter? Are there any assurances of the capability of the technician to correctly carry out the requested actions on behalf of a trained forensic investigator? Does the data returned from the Cloud system represent all of the stored data, including any inactive clusters that have not been overwritten? How accurate must the investigator instructions to the CSP be? Can the technician perform eDiscovery on behalf of the investigator or must subsequent warrants be issued in order to retrieve any additional information uncovered as relevant in the course of the analysis? To what extent is the CSP's system trusted to give a true and accurate image of the data stored? Is it a snapshot of static data or is it an image taken of live data and thus susceptible to "data smearing" much like a memory image in live forensics? In what way is the file metadata stored or created? Are timestamps generated uniformly and in a way that can be used to accurately perform forensic analysis and timeline extraction?

D. Live System Forensics

Traditional digital forensic techniques has generally been to "pull the plug, image afterwards" to prevent any contamination of the data during system shutdown, whether through deliberate anti-forensic acts or through automated system scripts that attempt to end a session tidily. More recently, the option of live forensic investigation has become almost a necessity as investigators find themselves trying to deal with encrypted hard drives or just encrypted containers often with a decryption key only residing in the RAM of the system being investigated. Live forensics allows the investigator to extract more metadata and more volatile data than post mortem analysis but it requires direct access to the system and may involve utilities that can cause `known` changes to the contents of memory or hard disks.

As described previously however, the cloud environment requires a different approach to a standard investigation. The constantly shifting pool of resources can present issues of its own outside of the challenges presented by the distributed nature of the data. If RAM is extracted from a VM, will the act of extraction cause any degradation to performance in the other guest systems on that segment of the cloud? If system storage is imaged what are the safeguards against infringing another Client's privacy by capturing re-allocated

[1] seedboxco.net

blocks in the process of being "zeroed". How much of the RAM is considered shared between tenants? Can network traffic be easily segregated to allow accurate and selective packet sniffing?

V. Existing Techniques for Cloud Forensics

Cloud Forensics is defined as [16] *"the application of digital forensics science in cloud computing environments. Technically, it consists of a hybrid forensic approach towards the generation of digital evidence. Organizationally, it involves interactions among cloud actors for the purpose of facilitating both internal and external investigations. Legally it often implies multi-jurisdictional and multi-tenant situations"*.

Currently cloud evidence gathering involves first identifying where the data is stored and seizing the relevant storage hardware to cloning the virtual machine, and subsequently performing "traditional" analysis on the captured data [17]. However, cloud systems can play host to multiple networks of IaaS/PaaS tenants or store data archives of several organisations. This can potentially lead to a huge amount of data to be processed which can in turn lead to massive slowdowns in productivity and backlogs in investigations an eventuality addressed by Quick and Choo in their 2014 paper [18]. One solution proposed is the use of reduction of the evidence to form a parallel process through the removal of known irrelevant data and deduplication. This parallel would deal with a lower volume of data and attempt to answer less ambiguous questions and so act as a form of triage for the investigative process as a whole. Only if the triage turned up a potentially suspect activity would the full dataset be investigated. This process of reduction presents a new set of challenges such as the method of collection, the process of reduction itself to ensure as little loss of detail as possible and what data mining technique can be used that requires a little inference as possible.

The collection of data from a cloud instance has been defined as requiring third party co-operation and possibly specialised software provided by the CSP through an API. The trust of this API is a justifiable concern as it is the only way to pull all of the disparate data caches together that form a VM guest instance on the cloud, whether that be an entire OS, a storage device or an instance of a software application. In his 2011 paper [19], Birk notes that as of 2009, very little research had been performed into the implication of performing forensics in the cloud. In the course of his analysis, Birk suggests that a scheme along the same lines as that suggested in 2007 by Juels and Kaliski could be used to allow a client the ability to prove that a file hosted on a cloud server could be retrieved and would be a forensically sound copy of the original. Birk goes on to suggest that investigations performed in an IaaS environment can potentially retrieve more metadata to support findings that those performed on SaaS and PaaS. This of course all depends on the level of logging the CSP makes available to the client and the degree of visibility of the underlying system allowed.

In general, an accepted technique for forensic examination of a cloud-based system is to perform a snapshot of the suspect system and use that snapshot as the source of evidence, potentially mitigating the need for any system downtime during the investigative process.

A. Virtual Machine Forensics

A Cloud is described as a collection of services dynamically allocated to a subscribing client and presented as an end product or service with no exposure of the underlying mechanics. Currently, the most efficient way to manage this allocation is through the use of virtualisation. Virtualisation involves inserting a layer of abstraction between the "bare metal" and the OS the client sees (whether that is an application running on an underlying OS in the form of PaaS or a collection of systems that the subscriber can arrange and link (IaaS)). This layer of abstraction is provided by a `hypervisor` which in conjunction with a `Virtual Machine Manager` (VMM) handles all low level operations on behalf of the guest operating system including the routing of all network traffic, allocation of disk space and aggregation of pagefiles to ensure efficiency in the utilisation of resources. Performing forensics in such an environment can be a challenging task [20] with its own set of restrictions and limitations, not least of which is the issue of shared resources.

Another, less obvious, issue arises when the investigator does not have access to the guest OS. Perhaps the system suspected of criminal activity is encrypted in some fashion. If the system was provisioned by the CSP but the end-user OS was configured by the Client, then, barring some form of backdoor built into all systems on that cloud platform (which would in itself be a security risk and not something a Client would usually agree to as part of an SLA) there is no way to read the data contained within the virtual instance. Anything that exists independent of the OS would be accessible, such as RAM and network connectivity (but not traffic content unless sent in an unencrypted format). In such a case the investigator may be able to use a method known as Virtual machine Introspection where the VMM is used to provide a restricted mode of input and output to and from the guest OS. One use of this VMM based I/O path is to perform kernel injection [21] which can trick the guest OS into executing a command. In a forensic investigation that command may be to open access on a port, start a service or even provide information about its current state and contents.

B. Private Cloud Forensics

A digital forensic investigation involving a private cloud is not as restricted in its methodology as those involving public, hybrid or community cloud deployments. The owner of the cloud is also the subscriber and so co-operation of one means co-operation of the other which can be voluntary or through court order. In addition, depending on the size of the organisation owning the cloud, the distribution across datastores is likely to be much less dispersed and may even be restricted to two or three sites (a primary site, a secondary site

for high availability and disaster recovery/business continuity and a final, redundant site for backup and emergency purposes) and these sites may or may not be housed in close proximity.

One such configuration is described by Martini and Choo [22], where they analyse the forensic process of investigating ownCloud, an open source cloud system intended to be deployed as a private StaaS. In their analysis they modify the accepted forensic model to include a stage where the client used to access the ownCloud system is used to enumerate the contents of the StaaS share provided and access to the cloud storage service is managed through a controlled utilisation of the client under investigation. Enumeration is performed using the client-side Sync and file management metadata and a time line can be constructed from the creation/modification and sync timestamps logged. In addition, evidence can be gathered from cached files as well as any logs or artefacts of authentication with the storage service itself (DNS lookups, URL history, cookies, certificates etc.). For the server itself, despite the cloud moniker, it should be approached as any standard hypervisor would be handled. If small enough, a forensic copy should be taken of the entire server, otherwise log files should be exported form the VMM and hypervisor and suspect systems should be either cloned to an external drive before being suspended or a snapshot taken and then exported as an image for post mortem analysis. In the case of ownCloud, the metadata to be extracted from the hypervisor and VMM is stored as either SQLite or MySQL databases and should be collected using the appropriate forensic utility.

C. Storage as a Service (StaaS Forensics)

While the most notable private solution, ownCloud is not the only StaaS system available and mobile access to consistent data or collaborative access to resources has spurred the availability of cloud-based StaaS solutions such as Dropbox [23] where files can be stored and shared with other users via an application or web-based console. Mulazzani also identifies a feature he terms "online slack space" which is similar to physical drive slack space where data resides in the unused portion of an individual block on a hard drive but in this case, data is stored in chunks on a Dropbox system without attribution to an owning system. While this allows for unlimited storage without decreasing the amount of allocated space, it is unclear what the security implications for this slack space storage are. Without an owner are the files protected by any of Dropbox's access controls? Will they be collected as unallocated blocks and destroyed as part of garbage collection on the system? Mulazzani's testing allowed retrieval of the slack space files after a period of four weeks.

Chung et al. [24] take this analysis further in their paper where they divide StaaS into three basic categories distinguished by the type of files that can be stored. In each of these categories they choose a service provider that fits the description and then perform a mock forensic analysis using a modified procedure that operates across all three categories. The main adaptation from the standard forensic methodology is that instead of limiting the scope of an investigation to each workstation or device taken isolation from one another, each device available belonging to a suspect must be compared. Then the evidence is gathered and merged into a single timeline of events to gain an accurate depiction of the activities in question.

Just as the cloud is ever changing, so too software. Developers change and alter its structure to provide additional functionality, close identified security holes or just adapt to changing back-end and client-side technological shifts (such as a new OS or platform to be catered for). In 2013, just one year after Chung, Quick and Choo [25] performed a follow up investigation on Dropbox, Google Drive and SkyDrive (recently renamed to OneDrive). In their process, Quick and Choo present the forensic artefacts left on the local system as part of installation or usage of the StaaS facility. In Dropbox, it was noted that the `.db` files identified by Chung as being of high value to an investigator as they contained historical synchronisation logs, had changed in format (from `.db` to `.dbx`) and were encrypted. As with Chung, Quick and Choo use a controlled execution of the service web-based front-end or client based application to enumerate the contents of the datastore.

VI. Cloud Facilitated Forensics

Employing cloud computing resources and computational power by law enforcement and digital investigators towards the analysis, indexing and storage of digital evidence can greatly expedite the forensic process. Some of the advantages of employing cloud-based forensic investigation include:

- A capacity to automatically triage evidence prior to full analysis [26].
- Reduced data storage requirements.
- An ability to quickly whitelist or blacklist files across numerous investigations.
- Focusing the investigation onto pertinent machines in an institution at an early point in the process.

A. Evidence Handling

Affordable, redundant cloud-based storage seems like an ideal place to store digital evidence. Employing data deduplication techniques (similar to those used by Storage as a Service providers, such as Dropbox) can greatly reduce the volume of information that is required to be stored. In this scenario, a single copy of each file is stored on the cloud and a hard drive image is recompiled from composite files when needed. In 2009, Watkins et al. coined the term "analytically sound" to refer to such reconstituted hard drive images containing the pertinent information to an investigation, unique to that suspect disk's imaging process. These images were created using an evidence acquisition tool called Teleporter [27], and while the reconstituted image from this system was found to be sufficient for forensic analysis, it is not a forensically sound bit-by-bit copy of the original source. On a file level, any hash values would match, but the hard drive image in its entirety would not match the original. Evidence deduplication (based on a software reference database, such as the NSRL database)

can greatly improve the imaging of enterprise machines by avoiding the waste of resources for handling harmless files [28], [29].

B. Cloud Facilitated Forensics of the Cloud

Performing cloud-to-cloud imaging or instance mirroring is the most performant option for forensics evidence acquisition, assuming an investigation instance can be launched on the same cloud platform as the suspect virtual machine. In 2014, Thethi and Keane found that imaging a cloud instance to another VM take the minimum amount of time in their cloud image acquisition performance testing work [30].

In this scenario, the instance itself, its associated storage and its network traffic could be continuously monitored (assuming relevant granted warrants) in order to gather information about the attacks or crimes it is aiding or where the sensitive information it is gathering is being transmitted. In order for this option to be viable, cooperation is again required on behalf of the CSP to facilitate such investigation.

VII. Conclusion

While the usage of cloud systems is ever increasing, the volume and global distribution of data potentially relevant to each digital forensic case. Cloud computing provides a significant technical and legal challenges to digital investigators and the solutions to many of these problems are still in their infancy. Many of the existing proposed solutions to these problems rely on cooperation of the CSPs. The CSPs are capable of providing tools and systems for law enforcement through forensic readiness, i.e., data acquisition methods, comprehensive log management, secure and reliable data provenance, etc. The provisioning of these systems may not have a benefit to the bottom-line for the CSP, but should remain a priority to provide from an ethical and legal standpoint. In the meantime, there is much work to be conducted by third-party investigators to enable forensic investigation to proceed despite current cooperative limitations.

A. Areas for Future Research

While cloud computing has become prevalent as a scalable solution for many businesses in recent years, there remains a number of challenges and areas for future research with regards to the investigation of these services [31].

Areas for future research in the area include:

- CSP Forensic API – CSP cooperation in the lawful forensic evidence acquisition and analysis of cloud data could provide efficient access law enforcement and digital investigators through specific APIs [17]. This avenue pushes the accountability and preparedness for forensic investigation to the cloud providers [32], [33], [34].
- Data Provenance in the Cloud – Due to the volatility and reuse frequency of cloud resources, it can often prove difficult, if not impossible to identify the origin of a pertinent piece of evidence. The development of a comprehensive secure provenance system for cloud resources will be crucial for future investigations.
- Comprehensive Log Management – Due to the multi-layer approach to the provisioning of many cloud-based services, the resolution and preservation of these logs is critical to building a timeline of events. A log management solution could provide instant access to the logs relevant to any interesting event that occurs across the cloud platform [35].
- Internet of Things (IoT) – The interaction between IoT devices and cloud computing back-ends will become more prevalent in coming years. The aggregation and processing of data collected from this devices will likely become a wealthy source of digital evidence information in future investigations [36].

References

[1] P. Balboni and E. Pelino, "Law enforcement agencies' activities in the cloud environment: a european legal perspective," *Information & Communications Technology Law*, vol. 22, no. 2, pp. 165–190, 2013.

[2] P. Mell and T. Grance, "The nist definition of cloud computing," 2011.

[3] S. A. Almulla, Y. Iraqi, and A. Jones, "A state-of-the-art review of cloud forensics," *Journal of Digital Forensics, Security and Law*, vol. 9, no. 4, pp. 7–28, 2014.

[4] N. Samet, A. Ben Letaifa, M. Hamdi, and S. Tabbane, "Forensic investigation in mobile cloud environment," in *Networks, Computers and Communications, The 2014 International Symposium on*. IEEE, 2014, pp. 1–5.

[5] Y. Xing and Y. Zhan, "Virtualization and cloud computing," in *Future Wireless Networks and Information Systems*, ser. Lecture Notes in Electrical Engineering, Y. Zhang, Ed. Springer Berlin Heidelberg, 2012, vol. 143, pp. 305–312. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-27323-0_39

[6] S. A. Baset, "Cloud slas: Present and future," *SIGOPS Oper. Syst. Rev.*, vol. 46, no. 2, pp. 57–66, Jul. 2012. [Online]. Available: http://doi.acm.org/10.1145/2331576.2331586

[7] R. A. Popa, J. R. Lorch, D. Molnar, H. J. Wang, and L. Zhuang, "Enabling security in cloud storage slas with cloudproof." in *USENIX Annual Technical Conference*, vol. 242, 2011.

[8] R. Lu, X. Lin, X. Liang, and X. S. Shen, "Secure provenance: the essential of bread and butter of data forensics in cloud computing," in *Proceedings of the 5th ACM Symposium on Information, Computer and Communications Security*. ACM, 2010, pp. 282–292.

[9] S. Zawoad, A. K. Dutta, and R. Hasan, "Seclaas: secure logging-as-a-service for cloud forensics," in *Proceedings of the 8th ACM SIGSAC symposium on Information, computer and communications security*. ACM, 2013, pp. 219–230.

[10] S. L. Garfinkel, "Digital forensics research: The next 10 years," *digital investigation*, vol. 7, pp. S64–S73, 2010.

[11] B. Grobauer and T. Schreck, "Towards incident handling in the cloud: challenges and approaches," in *Proceedings of the 2010 ACM workshop on Cloud computing security workshop*. ACM, 2010, pp. 77–86.

[12] R. W. Group, "The nist cloud computing roadmap 2013," 2013.

[13] H.-Y. Chen, "Cloud crime to traditional digital forensic legal and technical challenges and countermeasures," in *Advanced Research and Technology in Industry Applications (WARTIA), 2014 IEEE Workshop on*. IEEE, 2014, pp. 990–994.

[14] N. C. C. F. S. W. Group, "Nist cloud computing forensic science challenges," 2014.

[15] M. Scanlon, J. Farina, N.-A. Le Khac, and M.-T. Kechadi, "Leveraging Decentralisation to Extend the Digital Evidence Acquisition Window: Case Study on BitTorrent Sync," pp. 85–99, September 2014.

[16] K. Ruan, J. Carthy, T. Kechadi, and I. Baggili, "Cloud forensics definitions and critical criteria for cloud forensic capability: An overview of survey results," *Digital Investigation*, vol. 10, no. 1, pp. 34–43, 2013.

[17] Yannikos, "Cross-border cloud investigations," *Dagstuhl Reports*, vol. 3, no. 11, pp. 193–208, 2013.

[18] D. Quick and K.-K. R. Choo, "Impacts of increasing volume of digital forensic data: A survey and future research challenges," *Digital Investigation*, vol. 11, no. 4, pp. 273–294, 2014.

[19] D. Birk and C. Wegener, "Technical issues of forensic investigations in cloud computing environments," in *Systematic Approaches to Digital Forensic Engineering (SADFE), 2011 IEEE Sixth International Workshop on*. IEEE, 2011, pp. 1–10.

[20] B. Hay and K. Nance, "Forensics examination of volatile system data using virtual introspection," *ACM SIGOPS Operating Systems Review*, vol. 42, no. 3, pp. 74–82, 2008.

[21] P. Tobin and T. Kechadi, "Virtual machine forensics by means of introspection and kernel code injection," in *Proceedings of the 9th International Conference on Cyber Warfare & Security: ICCWS 2014*. Academic Conferences Limited, 2014, p. 294.

[22] B. Martini and K.-K. R. Choo, "Cloud storage forensics: owncloud as a case study," *Digital Investigation*, vol. 10, no. 4, pp. 287–299, 2013.

[23] M. Mulazzani, S. Schrittwieser, M. Leithner, M. Huber, and E. Weippl, "Dark clouds on the horizon: Using cloud storage as attack vector and online slack space." in *USENIX Security Symposium*, 2011.

[24] H. Chung, J. Park, S. Lee, and C. Kang, "Digital forensic investigation of cloud storage services," *Digital investigation*, vol. 9, no. 2, pp. 81–95, 2012.

[25] D. Quick and K.-K. R. Choo, "Forensic collection of cloud storage data: Does the act of collection result in changes to the data or its metadata?" *Digital Investigation*, vol. 10, no. 3, pp. 266–277, 2013.

[26] ——, "Data reduction and data mining framework for digital forensic evidence: storage, intelligence, review and archive," *Trends & Issues in Crime and Criminal Justice*, vol. 480, pp. 1–11, 2014.

[27] K. Watkins, M. McWhorte, J. Long, and B. Hill, "Teleporter: An analytically and forensically sound duplicate transfer system," *digital investigation*, vol. 6, pp. S43–S47, 2009.

[28] F. Cruz, A. Moser, and M. Cohen, "A scalable file based data store for forensic analysis," *Digital Investigation*, vol. 12, pp. S90–S101, 2015.

[29] N. C. Rowe, "Identifying forensically uninteresting files using a large corpus," in *Digital Forensics and Cyber Crime*. Springer, 2014, pp. 86–101.

[30] N. Thethi and A. Keane, "Digital forensics investigations in the cloud," in *Advance Computing Conference (IACC), 2014 IEEE International*. IEEE, 2014, pp. 1475–1480.

[31] S. Simou, C. Kalloniatis, E. Kavakli, and S. Gritzalis, "Cloud forensics solutions: A review," in *Advanced Information Systems Engineering Workshops*, ser. Lecture Notes in Business Information Processing, L. Iliadis, M. Papazoglou, and K. Pohl, Eds. Springer International Publishing, 2014, vol. 178, pp. 299–309. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-07869-4_28

[32] A. Haeberlen, "A case for the accountable cloud," *ACM SIGOPS Operating Systems Review*, vol. 44, no. 2, pp. 52–57, 2010.

[33] L. De Marco, M.-T. Kechadi, and F. Ferrucci, "Cloud forensic readiness: Foundations," in *Digital Forensics and Cyber Crime*. Springer, 2014, pp. 237–244.

[34] Z. Reichert, K. Richards, and K. Yoshigoe, "Automated forensic data acquisition in the cloud," in *Mobile Ad Hoc and Sensor Systems (MASS), 2014 IEEE 11th International Conference on*. IEEE, 2014, pp. 725–730.

[35] S. Zawoad and R. Hasan, "Cloud forensics: a meta-study of challenges, approaches, and open problems," *arXiv preprint arXiv:1302.6312*, 2013.

[36] R. Hegarty, D. Lamb, and A. Attwood, "Digital evidence challenges in the internet of things," in *Proceedings of the Tenth International Network Conference (INC 2014)*. Lulu. com, 2014, p. 163.