Forensic-as-a-Service (FaaS) in the Cloud State-of-the-Art

Avinash Srinivasan
Computer and Information Sciences
Temple University, Philadelphia, PA
avinash@temple.edu

Frank Ferrese
Electrical and Computer Engineering
Temple University, Philadelphia, PA
Philadelphia, PA
ferresef@temple.edu

Abstract

The need for digital forensic analysis in response to the unprecedented growth in the number of cases involving and depending on electronic data is at an all-time high. Rapid evolution of technology has further complicated matters necessitating the acquisition and analysis of digital evidence from a wide variety of media. Current digital forensic analysis capabilities utilizing standalone forensics workstations are arduously time-consuming and have been long surpassed by the ever-growing case backlog.

However, fundamental advances in the computing and communications industry has has catalyzed the transformation of *cloud computing* from a mere plausibility to a hard-reality and a survival necessity, especially for small and medium business enterprises. It has fueled numerous business opportunities in service industry as well as innovation across verticals that were previously in the realm of beyond available computing resources. It is now time to embrace the dawn of such evolution to develop innovative solutions to address the ever-growing and seemingly unsurmountable challenge faced by the digital forensics community.

Keywords: Cloud computing, cyber crime, digital forensics, distributed computing, electronic crime, electronic evidence, forensics investigation, investigation, integrity, parallel processing.

16.1 Introduction

Advances and fundamental changes in the computing and communications industry have resulted in significant challenges to current *Digital Forensics Analysis* practices, policies, and regulations. Consequently, the forensics analysis process is suffering from significant roadblocks not only from unclear cyber-laws and regulations, but also as a result of significant technology challenges. Integrity is the key requirement in the forensics analysis process. To further complicate matters, computer forensics analysis is fundamentally a serial process. Therefore, inherent scalability challenges exist. Most importantly, the ability to withstand the Daubert test during the trial is pivotal to designing parallel and distributed forensics analysis tools. In light of the above web of challenges, case backlogs are growing at an increasing rate. As noted in [Hitchcock et al., 2016], backlog is commonly in the order of 6-18 months, but can reach significantly higher numbers in some jurisdictions.

One instance of a key paradigm shift in the computing industry is the advent of *cloud computing*. In the recent years, *cloud computing* capabilities have advanced significantly and evolved from a mere plausible concept to hard reality of survival for many industries. It has brought along numerous business opportunities and everyone, from start-ups and small industries to Fortune-100, is embracing cloud computing, perhaps each from different viewpoints and varying business needs. Some of the attractive benefits of cloud computing include reduced in-house infrastructure burden, minimized maintenance and updates pressure, and the ability to quickly scale as computing needs increase. A 2013 Gartner report predicts that the cloud-based security services market, which includes secure email or web gateways, identity and access management (IAM), remote vulnerability assessment, security information and event management will surpass $4 billion by 2017 [Messmer, 2013].

16.1.1 Current State of Cloud Computing

Cloud computing is undoubtedly one of the most significant technology advances of the 21st century computing technology. The dawn of the cloud computing paradigm had three service delivery models, viz – *Software-as-a-Service (SaaS)*, *Platform-as-a-Service (PaaS)*, and *Infrastructure-as-a-Service (IaaS)*. However, innovation and advancement fueled by growing consumer and business needs led to the birth of numerous other delivery models such as – *Scanning-as-a-Service* [Gionta et al., 2014] and *Monitoring-as-a-Service* [Alhamazani et al., 2015].

On the flip-side, the cloud computing platform presents some very serious security and privacy concerns. The vast resource pool it offers has been and continues to be exploited by malicious actors. An adversary can easily exploit pool the resources in real-time for malicious reasons. This situation has transformed matters from bad to worse for the *Law Enforcement* (LE) and *Intelligence Community* (IC). Some of potential threats from the cloud computing platform can be evidenced from services such as Cybercrime-as-a-Service (CaaS) [Robinson, 2016]; Malware-as-a-Service [Drozhzhin, 2016]; Attacks-as-a-Service [Lemos, 2010]; Crimeware-as-a-Service (CaaS) [Sood and Enbody, 2013]; and Exploit-as-a-Service [Grier et al., 2012]. Today's cloud computing architectures, though very popular, are not designed to meet some of the stringent digital forensics requirements electronic evidence. The most important requirements that are impacted by cloud computing are chain-of-custody and data-provenance.

Numerous works have focused in this area including [Taylor et al., 2011, Birk and Wegener, 2011, Marty, 2011, Reilly et al., 2010, Sibiya et al., 2012, Zawoad and Hasan, 2012, Zawoad et al., 2013, Zawoad and Hasan, 2013b, Grispos et al., 2013, Dykstra and Sherman, 2012]. Of particular relevance is the work of Zawoad and Hasan [Zawoad and Hasan, 2013a], in which they note that many of the assumptions of digital forensics with regards to tools and techniques are not valid in cloud computing. In [Chen et al., 2013], the authors evaluate the implementation of a cloud-based security center for network security forensic analysis for processing stored network traffic using cloud computing platforms to find the malicious attacks.

16.1.2 What is this chapter about?

The primary focus of this chapter is the second category discussed below. This chapter is about solving the ever-increasing number of both criminal and civil cases with electronic evidence, increasing data and storage device sizes, devices that get connected to the IoT that have been and continue to be foot soldiers for geographically remote cyber criminals and nation states. Some in the forensics community – LE and Intel agencies, researchers, and practitioners – have turned toward parallel and distributed computing paradigms in the hopes of overcoming the seemingly unsurmountable

case backlog. One specific direction of interest is the cloud computing. It is now clear that utilization of cloud resources to accelerate the turn-around times for forensics investigations is inevitable and the its adoption on massive scales in both imminent and impending. Some of the early works to this aim include [Wei, 2004, Richard III and Roussev, 2006a, Richard III and Roussev, 2006b, Beebe and Clark, 2007, Liebrock et al., 2007, Marziale et al., 2007, Ayers, 2009, Reilly et al., 2010, Roussev et al., 2009].

16.1.3 Chapter Road-map

The remainder of this chapter is organized as follows. In section 16.2 we provide discussions on relevant background and present necessary preliminaries of this chapter. Then, in section 17, we review all existing state-of-the-art works focusing on parallel and distributed digital forensics analysis followed by discussions on the limitations in these works in section 17.2. In section 17.3 we present some of the key requirements to offering cloud-based FaaS. Finally, in section 17.4 we conclude the chapter with future research directions.

16.2 Background and Motivation

16.2.1 Limitations of Traditional Computer Forensics – Now & Forever

Today, it is not uncommon for laptops and desktops to be equipped with terabyte-sized storage. Similarly, in retrospect to a decade ago, digital forensics analysts today not only deal with significantly larger average disk size, but also with an extremely large variety of devices. Consequently, the amount of data that needs to be processed can run into tens of terabytes. Adding further to this problem is the number of cases today that require computer forensics analysis. On the other hand, as witnessed in recent crimes, attackers' level of sophistication has significantly advanced from the days of *Rabbit Virus "fork bomb"* and *Morris Worm* [M Chen and Robert, 2004] to state-of-the-art *Petya* and *Mirai* [Report,]. Even widely used commercial forensics suites such as EnCase[1] and FTK[2], are not keeping pace with increased complexity and data volumes of modern investigations. The growing burden on computer forensics analysts is evident from the reports published by FBI *Regional Computer Forensics Laboratories* (RCFLs) and *Computer Analysis Response Team* (CART). According to the 2010 RCFL annual report [RCFL, 2011], a total of $6,564$ examinations were conducted requiring processing of $3,086$TB data, with an average case size of 0.4TB. In its 2011 annual report [RCFL, 2012], the RCFL reported a total of $7,629$ examinations by analyzing $4,263$TB of data. During fiscal year 2012, the FBI CART supported nearly $10,400$ investigations, with over $13,300$ computer forensic examinations by processing data volumes in excess of $10,500$ TB [FBI-CART, 2013].

Numerous works in the recent years have tested the limits of traditional computer forensics tools and techniques to deal with evolving technology. Conventional wisdom may seem like computer systems should make investigations much faster simply by virtue of being able to perform billions of operations per second. In reality, however, the ever-increasing drive sizes necessitate significant (pre)processing times that far out-weighing the benefits of "billions of operations per second."

Limitations of first generation computer forensic tools are presented in [Ayers, 2009] along with metrics for measuring efficacy and performance of good tools. The author further lays out a broad set of requirements for second-generation tools and presents a high-level work-in-progress design for a second generation computer forensic analysis system. The ambition is to implement and test the

[1] https://www.guidancesoftware.com/

[2] http://accessdata.com/

prototype using two different processing architectures – i) *Beowulf cluster*, and ii) *IBM BlueGene/L super computer*.

'Forensic Cloud" is a framework for a forensic index-based search application presented in [Lee and Hong, 2011]. While it takes a substantial effort to construct an index database, the authors argue that searching though the indexed database returns query response in a fraction of the time it would take for the same query without indexing. Later, in [Lee and Un, 2012], Lee and Un present a case study supporting forensic indexed search as a service along with a work-in-progress model.

In their experiments, they achieve significantly better performance ($\approx$ 56MB/s) when the target data to be processed is more than 56GB. When a 1TB drive is analyzed with *bi-grams*, their system takes $\approx$ 2hours. Their system can also retrieve results from compressed text document formats at an average of $\approx$ 25MB/sec for a single query. Processing this query against a 1.27TB target took the authors $\approx$ 13 hours. However, they argue that this performance indeed outperforms existing forensic bitwise search methods by a significant margin. Further, the authors note that forensic bitwise search methods take $\approx$ 18.5 hours to perform a single keyword search on a 1TB drive.

This conclusion is supported further in [Roussev and Richard III, 2004] where the authors argue that a large part of the processing time is the "think" time, i.e., time needed for the human investigator to analyze the data. While it may be possible for a system to accumulate experience and reduce this time through Machine Learning, they are confident that the processing time needed by the system to execute "investigator issued" queries are largely dependent on the quality of the construction of the query.

In summary, the limitations of current forensics tools and techniques are deep rooted in the following: i) data diversity and abstraction, ii) I/O and processing speed, iii) I/O intensive tasks, iv) lack of automation, v) inability to scale, and vi) potential open-source tools that aren't yet approved.

16.2.2 Potential of Looking up to the Cloud – Forensics-as-a-Service

The cyberspace is highly dynamic and will not cease to evolve in its *applications*, *sophistication*, and reach. Consequently, the LE community will continue to work against the odds making forensics analysis ever-more challenging. In [Marziale et al., 2007], authors present compelling real-world use-cases justifying the need for more advanced tools. Their use-cases clearly demonstrate the inadequate capacity of traditional forensics investigation tools executing on a single workstation.

The time has come for a paradigm shift in computer forensics analysis. We require an adaptive, widely available, and priority driven parallel and distributed computing architecture. While cloud is inherently a distributed computing paradigm, its resourcefulness as a parallel computing paradigm has also been established [Ekanayake and Fox, 2009]. This migration to the cloud is necessary to both clear current backlogs as well as make it manageable in the future.

The Advanced Forensic Format (AFF) was proposed as an alternative to proprietary disk image formats. AFF is an open and extensible format for storing images of hard disks and other kinds of storage devices [Stevens et al., 2006]. The authors also present AFFLIB, an open source library that implements AFF. This work also proposed *Advanced Imager* (AIMAGE), which is a new disk image acquisition program that compares favorably with existing alternatives. This was later redesigned as AFF4 with backward compatibility by the authors in [Cohen et al., 2009]. The redesigned AFF4 format built upon the well supported ZIP file format specification making it simple to implement. Furthermore, the AFF4 implementation has downward comparability with existing AFF files.

17 State-of-the-Art in Parallel & Distributed Forensics Analysis

17.1 GPU-based Distributed Forensics Analysis

In [Gao et al., 2004] the authors discuss user and software engineering requirements for on-the-spot digital forensics tools to overcome time consuming in-depth forensic examinations. They present their *Bluepipe* architecture (shown in figure 1) for on-the-spot investigation along with the remote forensics protocol that they have developed.

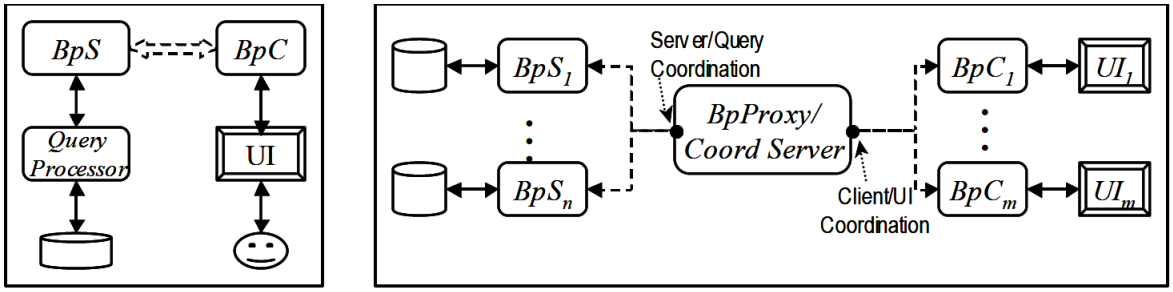

Figure 1: Bluepipe Architecture [Gao et al., 2004].

Feasibility of *Graphics Processing Units* GPUs for accelerating the traditional digital forensics analysis process is explored in [Marziale et al., 2007]. They note that the current generation of GPUs contains a large number of general purpose processors, in sharp contrast to previous generation designs, where special-purpose hardware units such as texture and vertex shaders were commonly used. This fact, combined with the prevalence of multi-core general purpose *Central Processing Units* CPUs in modern workstations suggests that performance-critical software such as digital forensics tools be "massively" threaded to take advantage of all available computational resources.

Results from a number of experiments that evaluate the effectiveness of offloading processing common to digital forensics tools to a GPU, using "massive" numbers of threads to parallelize the computation are presented in [Marziale et al., 2007]. These results are compared to speedups obtainable by simple threading schemes appropriate for multi-core CPUs which indicate that in many cases, the use of GPUs can substantially increase the performance of digital forensics tools.

In [Roussev and Richard III, 2004], the authors present the impact of evidence data size on analysis turn-around time. They evaluate the performance of the very popular commercial tool "*Forensics Tool Kit (FTK)*" by opening a case containing an old 6GB hard disk using the default options of the tool. During their study, FTK took approximately 2 hours to just open the case with the 6GB image. Using this time as the base line, and a conservative assumption that the processing time grows linearly as a function of size, authors conclude that it would take the state-of-the-art commercial tool approximately 60 hours to simply open a case with a 200GB disk image. However, in reality, when they tested their estimation on a 80GB image, it took FTK over 4 days (96+ hours) to just open the image. Therefore, there are indications that the tool does not scale linearly with increasing sizes of disk image.

Finally, in [Roussev and Richard III, 2004], authors weigh-in on the long-standing debate on whether to adopt a *Generic Distributed Framework* (GDF) for DDF purposes or to develop a more specialized solution. They conclude that a specialized solution is a better approach for the following reasons. First, they are more amenable to optimization for any specific purpose and, hence, can

achieve better performance with less overhead. Second, specialized solutions have minimized requirements of pre-installed infrastructure on all the machines. This enables regular users to run the system with ease and minimum administrative overhead. Finally, specialized solutions are better since GDFs have specialized programming interfaces requiring effort and experience for operator usage.

In summary, the conclusion of their work was that the fundamental resource constrains on Workstation class systems have been pushed to their processing and performance limits. Consequently, efforts focusing on task and resource optimizations will only result in marginal improvements, if any, on execution time.

17.1.1 XIRAF – XML Information Retrieval Approach to digital Forensics

In [Alink et al., 2006], the authors propose XIRAF, a prototype system for forensics analysis, which is an XML-based implementation aimed at managing and querying forensic traces extracted from digital evidence. XIRAF systematically applies forensic analysis tools to evidence files. Each forensics analysis tool that is used produces an output consisting of structured XML annotations capable of referring to regions in the corresponding evidence file. Furthermore, such annotations are stored in a persistent back-end such as an XML database (DB) that can be queried at a later time. For querying XIRAF's XML database, the authors have developed *XQuery* which is a custom query tool.

XIRAF's XML-based forensics analysis platform provides the forensic investigator with a powerful and feature rich query environment in which browsing, searching, and predefined query templates are all expressed as *XQuery* queries – XML DB queries. The authors address two key data processing problems that occur during the feature extraction and analysis phases of a computer system investigation:

1. **Evidence Quantity.** Modern computer systems are routinely equipped with hundreds of gigabytes of storage and a large investigation will often involve multiple systems, so the amount of data to process can run into terabytes. The amount of time available for processing this data is often limited (e.g., because of legal limitations). Also, the probability that a forensic investigator will miss important traces increases every day because there are simply too many objects to keep track of.

2. **Evidence Diversity.** A disk image contains a plethora of programs and file formats. This complicates processing and analysis and has led to a large number of special-purpose forensic analysis tools such as *browser history analyzers*, *file carvers*, *file-system analyzers*, *IRC and Chat analysis tools*, *registry analysis tools*, etc. While it is clear that the output of different tools can and should be combined in meaningful ways, it is difficult today to obtain an integrated view on the output from different tools. Furthermore, even if proprietary and commercial tools are approved and acceptable, it is highly unlikely that any forensic investigator would have the time and the knowledge to apply use any of the relevant tools to the case and evidence at hand. Hence the authors propose their XIRAF framework with the following key properties: i) clean separation between feature extraction and analysis; ii) single, XML-based output format for all forensic analysis tools; iii) XML DB for storing the XML annotations; and iv) custom query tool *XQuery* for querying analysis tools' XML output.

Since December 2010, the Netherlands Forensic Institute has been using XIRAF – a service-based approach for processing and investigating high volumes of seized digital material. Service-based XIRAF has over the years evolved significantly and become a standard for hundreds of criminal

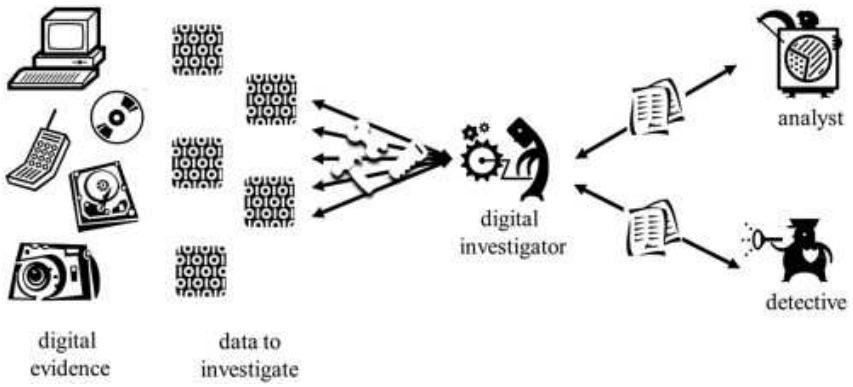

Figure 2: HANSKEN Architecture [Van Beek et al., 2015].

cases and over a thousand investigators, both in the Netherlands and other parts of the world. The authors note the impact of the XIRAF system and the paradigm shift it is causing, having processed over a petabyte of data with the XIRAF system.

XIRAF, which originally was conceived and initiated in 2006 as a scientific research project was primarily aimed at identifying and developing techniques for automating (parts of) the data analysis process for forensics investigations. XIRAF was never meant to be an operational system for processing large volumes of data, most definitely not data volumes in petabytes. Consequently, design considerations made during the development of XIRAF leave significant room for improvement.

17.1.2 HANSKEN – DFaaS Successor to XIRAF [Van Beek et al., 2015]

HANSKEN was well defined and designed from its inception and has a proof of concept (PoC) based on the new principles and ideas and a production version to replace XIRAF [Alink et al., 2006]. The forensic drivers behind the design and development of HANSKEN have been to provide a service that processes high volumes of digital material in a forensic context. In addition, it also provides easy and secure access to the processed results. The HANSKEN forensics framework is designed focusing on the following three drivers: i) minimization of the case lead time, ii) maximization of the trace coverage, and iii) specialization of people involved.

Processing of the seized material must be automated to provide the investigations team access to critical data. This impacts the way digital material is handled as noted in [Van Baar et al., 2014]. Furthermore, the results of this automated process must be made available to the investigation team directly and not to specialized digital investigators. To further speed up the investigation, analysts should be able to annotate or tag interesting traces such as those that need further analysis, or those that are not clear to the investigator who tagged it. Such annotation/tagging should be available to other analysts so that the case can be solved through collaboratively analysis.

The design of HANSKEN supports distributed extraction of traces from images. XIRAF, the precursor of HANSKEN, applies multiple tools to a forensic image on a single machine. This is iterative in nature and hence does not scale well. Most importantly, the design of XIRAF meant taking data to the tools, since tools are applied sequentially, with each tool having dedicated access to the image. To overcome this limitation of sequential processing, HANSKEN design was driven toward taking the tools to the data. HANSKEN uses distributed technology making it possible to process one forensic image using multiple machines. Consequently, as soon as the data is read from the image, it is kept in memory and all tools are applied. Once a trace is fully processed, the results are stored in a database so it can be queried while other traces are still being extracted.

This makes the first trace available in minutes, with more traces available for querying, mitigating idle time.

Another key feature of HANSKEN is its *data driven acquisition* such that analyst can start the process of extracting traces from a forensic image as soon as the first bits of a device are uploaded to the central system. To support this feature, authors have designed an image format that splits the image data in encrypted blocks. Such a format supports processing unordered blocks which makes it possible to implement *dynamic pipelining* where the extraction process influences the imaging process by asking for certain blocks of data to become available with priority.

17.1.3 MPI MapReduce (MMR) [Roussev et al., 2009]

The authors [Roussev et al., 2009] present three possible alternative approaches for augmenting forensics data processing in a fixed amount of time. The first is through the development of improved algorithms and tools for better and more efficient use of available machine resources. The second approach is the use of additional hardware resources to deploy additional machine resources. The third and the last alternative is by facilitating human collaboration leveraging human expertise in problem solving. All three approaches are mutually independent and support large-scale forensics in complementary ways.

The authors propose an open implementation of the *MapReduce* processing model that they call *MPI MapReduce* (MMR). The proposed MMR falls under the second category since it supports the use of additional hardware in the form of commodity distributed computational resources to speed-up forensic investigations.

MMR's performance has been evaluated through a proof-of-concept implementation leveraging two key technologies. The first is the *Phoenix Shared-Memory implementation* of MapReduce. The second is the *Message Passing Interface* (MPI) distributed communication standard. In summary, MMR [Roussev et al., 2009] MMR provides linear scaling for CPU-intensive processing and super-linear scaling for indexing-related workloads.

17.1.4 GRR Rapid Response Framework [Cohen et al., 2011]

GRR Rapid Response Framework (GRR), a new multi-platform, open source tool for enterprise forensic investigations was presented in [Cohen et al., 2011]. A key feature of GRR is its ability to support remote raw disk and memory access. GRR is designed to be scalable and it is a distributed approach for remote live access that is intended to be scalable. However, it is not a cloud-based solution, instead a live forensics tool geared towards preserving volatile evidence. Yet another remote access technique utilized is presented in [Cohen, 2005]. The advantage of is this technique is that the client side is very simple, while the server side performs the complex forensic analysis.

A key challenge to automate analysis is that it may require executing many sequential steps. Current solutions create a dedicated console process that waits for the client to complete each step before issuing the next step. This limits scalability as the server needs to allocate resources for each client and wait until the entire analysis is complete. In GRR, the authors use state serialization to suspend execution for analysis processes for each client. These serialized forms are then stored dormant on disk until the client responds. Consequently, this approach resolves the problem of resource drain imposed on servers. In GRR, such constructions are referred to as *Flows*. A *Flow* is simply a state machine with well-defined serialization points, where it is possible to suspend its execution.

The architecture of GRR addresses auditing and privacy issues by allowing for non-intrusive automated analysis with audited access to retrieved data. However, it strives to achieve a balance

between protecting access to user data and warranted forensically sound analysis. It also provides a secure and scalable platform to facilitate employ a variety of forensic analysis solutions. The authors support the usefulness and practicality of their proposed GRR through the following four case studies: i) Investigation of intellectual property leaks; ii) Isolation of targeted malware attack; iii) Discovery requests' compliance; and iv) Periodic snapshots of systems' states.

17.1.5 A scalable file based data store for forensic analysis [Cruz et al., 2015]

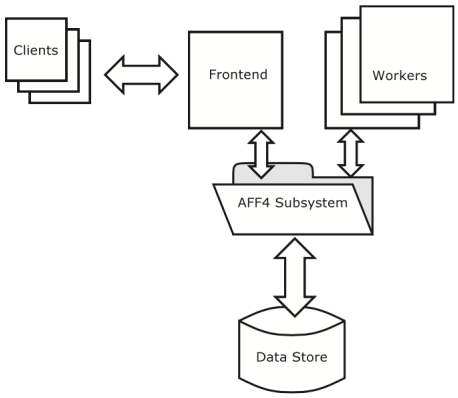

Figure 3: GRR Architecture [Cruz et al., 2015].

In [Cruz et al., 2015], the authors present the *GRR Rapid Response* (GRR) framework, which is a live forensic and incident response framework. The GRR architecture is constructed over the AFF4 subsystem [Cohen et al., 2009], which essentially implies that all data stored in the GRR data store are comprised of serialized AFF4 objects. In this work, the authors present a new data store back-end that can be used as a storage layer for the AFF4 Resolver. GRR's AFF4 Resolver stores AFF4 objects permanently inside a "*NoSQL*" data store enabling the application to only deal with high level objects. The proposed GRR's distributed data store partitions data into database files that can be accessed independently enabling scalable distributed forensic analysis. Furthermore, they discuss utilizing the software reference database *National Software Reference Library* (NSRL) in tandem with their distributed data store to avoid wasting resources when collecting/processing benign files. The following two functionalities must be implemented by the data store in order to support an AFF4 Resolver.

1. *Single object access* – simplifies the partitioning of data because operations never deal with multiple objects. GRR systems require synchronous operations to guarantee globally deterministic ordering.

2. *Support for synchronous and asynchronous operations* – synchronous operations will block until the data store returns the results, while asynchronous operations will be scheduled to be performed at some point in the future. Asynchronous operations improve program concurrency and provide a huge performance advantage, hence heavily used by GRR systems.

Originally, the SQLite data store provided by GRR exhibits two limitations: i) the capacity of each single worker degrades as new workers are added to the GRR system due to contention at the data store, which limits its *horizontal scaling*; and ii) since existing data stores rely on a central database server, increasing storage demands on a single server is only possible to a certain extent, which limits its *storage scaling*.

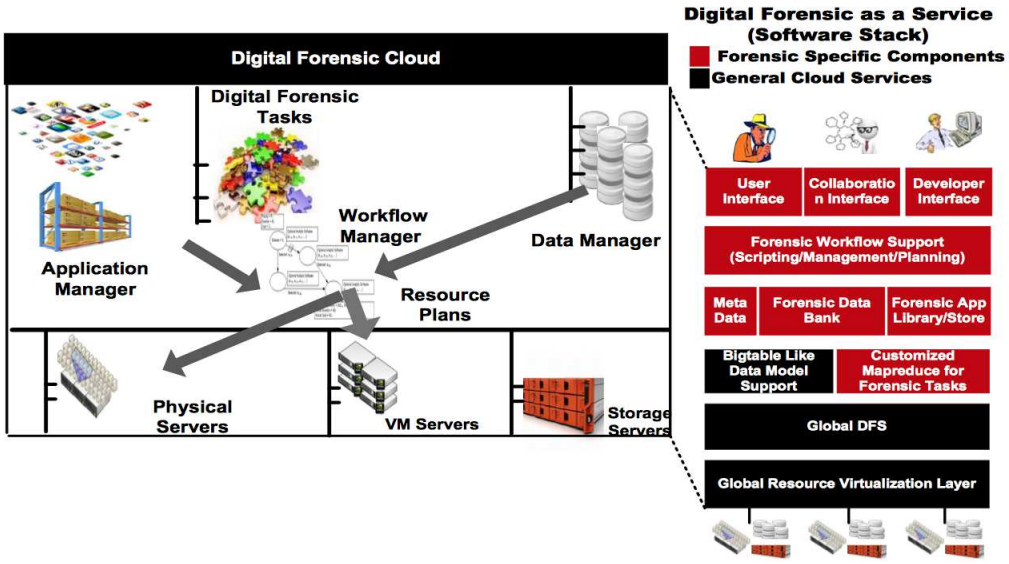

Figure 4: Digital Forensic-as-a-Service Software Stack [Wen et al., 2013].

According to the authors [Cruz et al., 2015], they reason that the above limitations are due to file lock contention at the central server. Therefore, they approach to resolve this problem by completely dividing the AFF4 name-space into independent storage files. This helps mitigate the file-lock contention problems. The benefits of their approach can be witnessed in the validation results in [Cruz et al., 2015].

17.1.6 Forensic-as-a-Service [Wen et al., 2013]

Wen et. al. [Wen et al., 2013] have proposed a domain specific cloud environment that can leverage the emerging trends of service-based computing, for supporting forensic applications. The proposed cloud-based forensics framework is specifically designed for dealing with large volume of forensic data. Furthermore, their approach also has the ability to enable the sharing of inter-operable forensic software and providing tools for forensic investigators to create and customize forensic data processing work-flows. Authors have conducted experiments using their forensic cloud framework using Amazon's Elastic Compute Cloud (EC2) service.

The experimental infrastructure is based on Hadoop 0.20 and HBase 0.20 and is managed by *Cloudear*[3]. For evaluations, the workloads are parallelized, and the results show that their approach can reduce forensic data analysis time considerably. They also argue that the overhead for the investigators to design and configure complex forensic work-flows is greatly minimized. Finally, they claim their proposed work-flow management solution can save up to 87% of analysis time in the tested scenarios.

17.1.7 Data De-duplication [Scanlon, 2016, Wolahan et al., 2016] driven Acceleration of Forensics Analysis

The authors [Scanlon, 2016, Wolahan et al., 2016] present a unique perspective to combat the digital forensic backlog. The proposed method explores a data de-duplication framework to eliminate

[3]http://www.cloudera/

redundancy in reacquisition, storage, and analysis of previously processed data. The primary objective of the authors in this case is to design a system that can alleviate some of the backlog by minimizing the duplicated efforts, while providing a number of enhancements to the functionality available with the traditional alternative. In [Wolahan et al., 2016], the authors explore alternative to the traditional evidence acquisition model through the leveraging of a forensic data deduplication system. This work also presents the advantages of a deduplicated approach along with some preliminary results of a prototype implementation.

17.2 Limitations in State-of-the-art Works

Digital forensic evidence acquisition speed is traditionally limited by two main factors: i) the *read speed* of the storage device being acquired; and ii) the *write speed* of the system the evidence is being acquired to. None of the above key works in distributed and parallel forensics analysis have addressed this issue. They are assuming that the data used are collected from different sources in a distributed way including using the cloud during acquisition. This is not a realistic assumption.

Typically, the first responders collect and image all of the evidences and then its uploaded onto the cloud. Unless multiple systems are being imaged, using cloud for acquiring evidence images does not yield better results due to system I/O limitations noted previously. Additionally, authors have not tested their framework on disk image without ground truth. Knowing the ground truth of an image and then processing it with tailored work-flow management will only yield good results. Therefore, the process efficiency and the speedup claim questionable.

In [Cohen et al., 2011], the proposed system provisions remote access to networked systems. However, the tool is specifically designed for remote live forensics of the networked systems. In [Wen et al., 2013], authors note that the data they use for experiments are collected from different sources in a distributed way using the cloud. They further note that the forensic data manager provides supports for uploading the evidence files to the cloud. However, the uploading time of evidence files is not considered when evaluating the performance of their framework. Therefore, speedup results they report does not truly reflect the actual speed up since uploading the evidence files is one of the most time-consuming steps in forensics analysis.

Another key area that has not been addressed is the difficulty in merging the analysis results from various tools into a single case report. Note that frameworks [Van Beek et al., 2015] that facilitate execution of various tools on the evidence images need a streamlined approach for consolidating the output into a meaningful analysis report.

17.3 Cloud-based Forensics-as-a-Service (FaaS)

17.3.1 Security and Privacy Requirements

FaaS service providers must assure all stakeholders, suspect, victim, judge, and jury, that their implementation and operations of forensics-as-a-service (FaaS) meet the regulatory standards for security and privacy of data and integrity requirements of the forensics processes. A FaaS service provider is expected to assure its stakeholders the three core security requirements with regards to case and evidence data security and privacy. The service provider must ensure that resource pooling in a multi-tenant environment does not risk the fundamental requirements of the security triad.

Confidentiality of case relevant information and evidence data is a key requirement. Service provider must ensure appropriate control mechanisms are in place to prevent accidental or intentional data disclosure, unauthorized access, or accidental/intentional data leaks either during or after the case analysis is complete. Furthermore, any potential confidentiality violations of users can

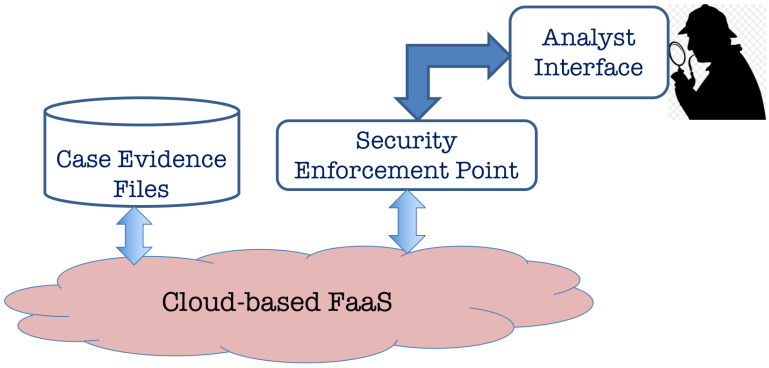

Figure 5: Security enforcement in FaaS.

potentially have a domino effect resulting in secondary violations such as *Health Insurance Portability and Accountability Act* (HIPAA) or *Family Educational Rights and Privacy Act* (FERPA). Similarly, third-party tools must also be thoroughly vetted to detect any potential data leaks.

The integrity of case relevant data is of even greater significance in the realm of computer forensics analysis. The FaaS provider mush have well established and tested integrity controls enforced to counter any potential risk of accidental or intentional alterations to case information and more importantly evidence data. Failure to implement strong integrity preserving security mechanisms can be catastrophic to the digital investigations with the potential of rendering the entire evidence data inadmissible.

Finally, the case information and evidence data should be available whenever authorized users need access. Though non-availability is not a critical security concern to the investigation, it can impact the indirectly due to down time resulting in delayed analysis. This can cascade to discovery of information that could warrant additional seizures, which may have been destroyed irreversibly. Also, at the completion of the analysis, there has to be proper procedures for backup and archiving to ensure availability of case relevant data in the future for (re)appeals or other legal purposes.

17.3.2 Regulatory and Legal Requirements

Compliance in the realm of Information Security is a fundamental requirement. A majority of enterprise forensics investigations involve non-compliant matters involving employees or the employer. Forensics investigation can span the whole spectrum of possibilities – from enterprise policy violations to insider threats, harassing emails to cyber stalking, robbery to vandalism, and suicide to homicide. Digital forensics investigations should comply with key regulations.

1. Strict control over cloud infrastructure and resources ensuring consistency in jurisdiction and applicable laws of the FaaS platform itself.

2. The FaaS platform and the entire process of analysis is monitored and logged at appropriate granularity enabling audit by a neutral and trusted third-party. The logs are themselves secured such that the neutral and trusted third-party auditing will not have access to any sensitive information such as PII during the course of the audit.

3. All methods, tools and techniques must be validated and approved by appropriate government authorities. One of the key approvals often comes from the NIST's *Computer Forensics Tool*

Testing (CFTT) program. Failure to prove the integrity and reproducibility of the process would render all efforts futile in court of law.

17.3.3 Design Requirements

Some of the key requirements for designing a parallel and distributed digital forensics toolkit framework are delineated below.

1. **Modular** – Since the entire forensics analysis is a complex process, a modular design facilitates a systematic breakdown of the complex process. Subsequently, tools and techniques can be developed for smaller tasks at level of granularity and abstraction that supports the case hypothesis. A modular design enables flexibility and extensibility, two key requirements to cope with evolving technology and threat landscape. Modular design enables rapid development of newer tools and their easy integration into the master tool framework.

2. **Scalable** – the architecture of the FaaS should be capable of scaling well with increasing numbers and sizes of cases and associated evidences. Increasing workload should not compromise the resource allocation and execution capabilities. A digital forensic analysis process is scalable if it can keep the average time per investigation constant in the face of growing target sizes and diversity [Roussev and Quates, 2012].

3. **Platform Independent** – FaaS should be able to handle forensics tools independent of the tools' needs for specific hardware/software platform. Furthermore, for the FaaS framework it should be possible to pool together machine resources of a group of investigators working on the same case to speed up the processing of critical evidence [Roussev and Richard III, 2004].

4. **Extensible** – Cloud-based FaaS framework should be devoid of vendor-locked functionality and capability expansions. This is a critical requirement for enhancing the FaaS relevance and capabilities to be current with evolving technology needs and case loads. Note that this is a standard software engineering requirement and it mandates that it should be easy to add new or replace existing functions [Roussev and Richard III, 2004].

17.3.4 Benefits of Provisioning Cloud-based Forensics-as-a-Service

By migrating the computer forensics analysis process to the cloud, the digital forensic science discipline will experience a broad spectrum of benefits. The first and foremost benefit would be more efficient utilization of limited manpower with required skills. This would also mean improved consistency in results from forensics analysis. Since cloud already offers metered services, migrating the forensics analysis process to the cloud will result in improved resource utilization while minimizing the cost. Since cloud as a computing platform is ubiquitous and widely accessible, it enables better inter-agency and intra-agency information and resource sharing. Furthermore, FaaS will offer consistent analysis platforms and resource allocations through established baseline. Finally, the most important benefit of FaaS would be provisioning accreditation and certification bodies convenient access to tools and processes for validation and certification.

17.4 Conclusion and Future Research Direction

Current trends in computing and communications technologies are putting vast amounts of disk storage and abundant bandwidth in the hands of ordinary computer users. These trends have long surpassed the capabilities of traditional workstation-based platforms for computer forensics. There

is plenty of evidence in existing body of works that address the limitations of current generation of tools and technologies from different perspectives. However, timely processing of digital data is still fundamental to computer forensics analysis. Consequently, large-scale distributed computing resources coupled with flexibility to customize the forensics processing performed is the critical.

There have been some initial attempts to leverage parallel and distributed computing paradigms to address a plethora of challenges faced by computer forensics analysts. In [Roussev et al., 2009], the authors have developed MPI MapReduce (MMR) as an alternative to Hadoop and demonstrated that the basic building blocks of many forensic tools can be efficiently realized using the MapReduce framework. Nonetheless, the true power of cloud computing is yet to be fully explored providing a ubiquitous Forensics-as-a-Service platform. The future for accelerating digital forensics analysis to keep pace with the ever-evolving technology and complexities in computer forensics analysis is inevitably in the direction of parallel and distributed computing. In particular, the ubiquitous and plentiful resource availability in the cloud is the most promising option to alleviate most of the problems currently faced, if not all.

References

[Alhamazani et al., 2015] Alhamazani, K., Ranjan, R., Jayaraman, P. P., Mitra, K., Liu, C., Rabhi, F., Georgakopoulos, D., and Wang, L. (2015). Cross-layer multi-cloud real-time application qos monitoring and benchmarking as-a-service framework. *arXiv preprint arXiv:1502.00206*.

[Alink et al., 2006] Alink, W., Bhoedjang, R., Boncz, P. A., and de Vries, A. P. (2006). Xiraf–xml-based indexing and querying for digital forensics. *digital investigation*, 3:50–58.

[Ayers, 2009] Ayers, D. (2009). A second generation computer forensic analysis system. *digital investigation*, 6:S34–S42.

[Beebe and Clark, 2007] Beebe, N. L. and Clark, J. G. (2007). Digital forensic text string searching: Improving information retrieval effectiveness by thematically clustering search results. *Digital investigation*, 4:49–54.

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

[Chen et al., 2013] Chen, Z., Han, F., Cao, J., Jiang, X., and Chen, S. (2013). Cloud computing-based forensic analysis for collaborative network security management system. *Tsinghua science and technology*, 18(1):40–50.

[Cohen, 2005] Cohen, M. (2005). Hooking io calls for multi-format image support. `http://www.sleuthkit.org/informer/sleuthkit-informer-19.txt`.

[Cohen et al., 2011] Cohen, M., Bilby, D., and Caronni, G. (2011). Distributed forensics and incident response in the enterprise. *digital investigation*, 8:S101–S110.

[Cohen et al., 2009] Cohen, M., Garfinkel, S., and Schatz, B. (2009). Extending the advanced forensic format to accommodate multiple data sources, logical evidence, arbitrary information and forensic workflow. *digital investigation*, 6:S57–S68.

[Cruz et al., 2015] Cruz, F., Moser, A., and Cohen, M. (2015). A scalable file based data store for forensic analysis. *Digital Investigation*, 12:S90–S101.

[Drozhzhin, 2016] Drozhzhin, A. (2016). Adwind malware-as-a-service hits more than 400,000 users globally. *Kaspersky Lab*.

[Dykstra and Sherman, 2012] Dykstra, J. and Sherman, A. T. (2012). Acquiring forensic evidence from infrastructure-as-a-service cloud computing: Exploring and evaluating tools, trust, and techniques. *Digital Investigation*, 9:S90–S98.

[Ekanayake and Fox, 2009] Ekanayake, J. and Fox, G. (2009). High performance parallel computing with clouds and cloud technologies. In *International Conference on Cloud Computing*, pages 20–38. Springer.

[FBI-CART, 2013] FBI-CART (2013). Piecing together digital evidence – the computer analysis response team. `http://www.fbi.gov/news/stories/2013/january/piecing-together-digital-evidence`.

[Gao et al., 2004] Gao, Y., Richard III, G. G., and Roussev, V. (2004). Bluepipe: A scalable architecture for on-the-spot digital forensics. *International Journal of Digital Evidence (IJDE)*, 3.

[Gionta et al., 2014] Gionta, J., Azab, A., Enck, W., Ning, P., and Zhang, X. (2014). Seer: practical memory virus scanning as a service. In *Proceedings of the 30th Annual Computer Security Applications Conference*, pages 186–195. ACM.

[Grier et al., 2012] Grier, C., Ballard, L., Caballero, J., Chachra, N., Dietrich, C. J., Levchenko, K., Mavrommatis, P., McCoy, D., Nappa, A., Pitsillidis, A., et al. (2012). Manufacturing compromise: the emergence of exploit-as-a-service. In *Proceedings of the 2012 ACM conference on Computer and communications security*, pages 821–832. ACM.

[Grispos et al., 2013] Grispos, G., Storer, T., and Glisson, W. B. (2013). Calm before the storm: the challenges of cloud. *Emerging Digital Forensics Applications for Crime Detection, Prevention, and Security*, 4:28–48.

[Hitchcock et al., 2016] Hitchcock, B., Le-Khac, N.-A., and Scanlon, M. (2016). Tiered forensic methodology model for digital field triage by non-digital evidence specialists. *Digital investigation*, 16:S75–S85.

[Lee and Hong, 2011] Lee, J. and Hong, D. (2011). Pervasive forensic analysis based on mobile cloud computing. In *Multimedia Information Networking and Security (MINES), 2011 Third International Conference on*, pages 572–576. IEEE.

[Lee and Un, 2012] Lee, J. and Un, S. (2012). Digital forensics as a service: A case study of forensic indexed search. In *2012 International Conference on ICT Convergence (ICTC)*, pages 499–503.

[Lemos, 2010] Lemos, R. (2010). Criminals 'go cloud' with attacks-as-a-service. Technical report, University of Zurich, Department of Informatics.

[Liebrock et al., 2007] Liebrock, L. M., Marrero, N., Burton, D. P., Prine, R., Cornelius, E., Shakamuri, M., and Urias, V. (2007). A preliminary design for digital forensics analysis of terabyte size data sets. In *Proceedings of the 2007 ACM symposium on Applied computing*, pages 190–191. ACM.

[M Chen and Robert, 2004] M Chen, T. and Robert, J.-M. (2004). The evolution of viruses and worms.

[Marty, 2011] Marty, R. (2011). Cloud application logging for forensics. In *Proceedings of the 2011 ACM Symposium on Applied Computing*, pages 178–184. ACM.

[Marziale et al., 2007] Marziale, L., Richard III, G. G., and Roussev, V. (2007). Massive threading: Using gpus to increase the performance of digital forensics tools. *digital investigation*, 4:73–81.

[Messmer, 2013] Messmer, E. (October 2013). Calm before the storm: The challenges of cloud computing in digital forensics. *Network World*.

[RCFL, 2011] RCFL (2011). Annual report for fiscal year 2010.

[RCFL, 2012] RCFL (2012). Annual report for fiscal year 2011.

[Reilly et al., 2010] Reilly, D., Wren, C., and Berry, T. (2010). Cloud computing: Forensic challenges for law enforcement. In *Internet Technology and Secured Transactions (ICITST), 2010 International Conference for*, pages 1–7. IEEE.

[Report,] Report, A. Q. . State of the internet / security. 4(2).

[Richard III and Roussev, 2006a] Richard III, G. G. and Roussev, V. (2006a). Digital forensics tools: the next generation. *Digital crime and forensic science in cyberspace*, pages 76–91.

[Richard III and Roussev, 2006b] Richard III, G. G. and Roussev, V. (2006b). Next-generation digital forensics. *Communications of the ACM*, 49(2):76–80.

[Robinson, 2016] Robinson, R. M. (2016). Cybercrime-as-a-service poses a growing challenge. *Security Intelligence*.

[Roussev and Quates, 2012] Roussev, V. and Quates, C. (2012). Content triage with similarity digests: The m57 case study. *Digital Investigation*, 9:S60–S68.

[Roussev and Richard III, 2004] Roussev, V. and Richard III, G. G. (2004). Breaking the performance wall: The case for distributed digital forensics. In *Proceedings of the 2004 Digital Forensics Research Workshop (DFRWS 2004*, volume 94.

[Roussev et al., 2009] Roussev, V., Wang, L., Richard, G., and Marziale, L. (2009). *A Cloud Computing Platform for Large-Scale Forensic Computing*, pages 201–214. Springer Berlin Heidelberg, Berlin, Heidelberg.

[Scanlon, 2016] Scanlon, M. (2016). Battling the digital forensic backlog through data deduplication. In *Innovative Computing Technology (INTECH), 2016 Sixth International Conference on*, pages 10–14. IEEE.

[Sibiya et al., 2012] Sibiya, G., Venter, H. S., and Fogwill, T. (2012). Digital forensic framework for a cloud environment.

[Sood and Enbody, 2013] Sood, A. K. and Enbody, R. J. (2013). Crimeware-as-a-servicea survey of commoditized crimeware in the underground market. *International Journal of Critical Infrastructure Protection*, 6(1):28–38.

[Stevens et al., 2006] Stevens, C., Malan, D., Garfinkel, S., Dubec, K.-A., and Pham, C. (2006). Advanced forensic format: An open, extensible format for disk imaging. International Federation for Information Processing.

[Taylor et al., 2011] Taylor, M., Haggerty, J., Gresty, D., and Lamb, D. (2011). Forensic investigation of cloud computing systems. *Network Security*, 2011(3):4–10.

[Van Baar et al., 2014] Van Baar, R., Van Beek, H., and van Eijk, E. (2014). Digital forensics as a service: A game changer. *Digital Investigation*, 11:S54–S62.

[Van Beek et al., 2015] Van Beek, H., Van Eijk, E., Van Baar, R., Ugen, M., Bodde, J., and Siemelink, A. (2015). Digital forensics as a service: Game on. *Digital Investigation*, 15:20–38.

[Wei, 2004] Wei, R. (2004). A framework of distributed agent-based network forensics system. *Proceedings of DFRWS*.

[Wen et al., 2013] Wen, Y., Man, X., Le, K., and Shi, W. (2013). Forensics-as-a-service (faas): computer forensic workflow management and processing using cloud. *Fifth International Conferences on Pervasive Patterns and Applications*, pages 1–7.

[Wolahan et al., 2016] Wolahan, H., Lorenzo, C. C., Bou-Harb, E., and Scanlon, M. (2016). Towards the leveraging of data deduplication to break the disk acquisition speed limit. In *2016 8th IFIP International Conference on New Technologies, Mobility and Security (NTMS)*, pages 1–5.

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

[Zawoad and Hasan, 2012] Zawoad, S. and Hasan, R. (2012). I have the proof: providing proofs of past data possession in cloud forensics. In *Cyber Security (CyberSecurity), 2012 International Conference on*, pages 75–82. IEEE.

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

[Zawoad and Hasan, 2013b] Zawoad, S. and Hasan, R. (2013b). Digital forensics in the cloud. Technical report, DTIC Document.