Towards the Leveraging of Data Deduplication to Break the Disk Acquisition Speed Limit

Hannah Wolahan[§], Claudio Chico Lorenzo[†], Elias Bou-Harb[‡], Mark Scanlon[*]

[*][§]School of Computer Science, University College Dublin, Ireland
[†]IES. Puerto de la Cruz, Spain
[‡]Cyber Threat Intelligence Laboratory, Florida Atlantic University, USA

Email: [*]mark.scanlon@ucd.ie, [§]hannah.wolahan@ucdconnect.ie, [†]claudiochicolorenzo@gmail.com, [‡]ebouharb@fau.edu

Abstract—Digital forensic evidence acquisition speed is traditionally limited by two main factors: the read speed of the storage device being investigated, i.e., the read speed of the disk, memory, remote storage, mobile device, etc.), and the write speed of the system used for storing the acquired data. Digital forensic investigators can somewhat mitigate the latter issue through the use of high-speed storage options, such as networked RAID storage, in the controlled environment of the forensic laboratory. However, traditionally, little can be done to improve the acquisition speed past its physical read speed from the target device itself. The protracted time taken for data acquisition wastes digital forensic experts' time, contributes to digital forensic investigation backlogs worldwide, and delays pertinent information from potentially influencing the direction of an investigation. In a remote acquisition scenario, a third contributing factor can also become a detriment to the overall acquisition time – typically the Internet upload speed of the acquisition system. This paper explores an alternative to the traditional evidence acquisition model through the leveraging of a forensic data deduplication system. The advantages that a deduplicated approach can provide over the current digital forensic evidence acquisition process are outlined and some preliminary results of a prototype implementation are discussed.

Index Terms—Digital Forensic Backlog; Data Deduplication; Evidence Acquisition

I. INTRODUCTION

One of the most significant issues facing the digital forensic communities around the world is attempting to deal with the sheer volume of cases requiring specialist analysis [4, 8]. A significant case backlog is commonplace in law enforcement agencies across the globe with legal proceedings often being heard with potentially pertinent evidence sitting in evidence storage lockers remaining unanalysed. This backlog is commonly in the order of 6-18 months, but can reach the order of years in some jurisdictions [2]. This delay in digital evidence processing can have a detrimental effect on the justice system [1]. A large number of factors can contribute to this backlog, and a number of these are outlined below:

- Expertise Availability – Law enforcement agencies have a limited amount of trained personnel to process the digital evidence in a forensically-sound, court-admissible manner.

- Variety of Devices and Data Types – The sheer diversity of devices types, storage media and remote sources of data greatly hinder digital investigative progress. With the advent of cloud-based storage of much of consumer level data, the traditionally simple device identification step of the digital forensic process has now become significantly more complex.

- Protracted Acquisition and Analysis Time – The acquisition and analysis phase are extremely time-heavy necessities of all digital forensics investigations. Acquiring evidence from physical or remote storage is limited by the read or upload speeds respectively.

While the first two factors can only be effectively tackled through training, education and expertise, this paper focuses on a technical solution capable of greatly expediting the last item in the above list. Through the leveraging of a deduplicated forensic data storage system, the acquisition and analysis phases of a typical investigation can be significantly expedited. Eliminating the unnecessary reacquisition and analysis of previously processed data can greatly expedite the current forensic process. The current digital forensic process typically starts with seizure of equipment from a suspect. This equipment is brought to an evidence store where it awaits further processing by expert personnel. Once the case reaches the top of the queue, the storage devices (hard drives, solid state drives, memory cards, external storage devices, etc.) are acquired using industry standard tools, e.g., EnCase or FTK Imager. This image (with a file size matching the original device) is then available for analysis on the forensic workstation or stored on a shared network storage device in the forensic laboratory.

A. Contribution of this Work

The contribution of this work can be summarised as:

- The design of a data deduplication system capable of the secure, verifiable acquisition of digital storage media is outlined.

- A method for the reconstruction of an entire disk image from a deduplicated data store.

- The results of an evidence acquisition of numerous hard drive images to a proof-of-concept implementation of the deduplication system are presented and evaluated.

II. RELATED WORK

A. Expediting the Digital Forensic Process

Current approaches to expedite the digital forensic investigative process focus on a range of aspects and techniques to speed up the analysis and contraband identification process. Most techniques to date broadly fall into the following approaches:

- Triage – Digital forensic field triage involves providing minimal training to front-line law enforcement in the identification, handling and seizure of digital forensic evidence. An extension to this methodology is to provide remote access to suspect devices to digital forensic experts to enable time sensitive analysis and investigation. Fast detection of pertinent evidence can prove crucial to the timely progression of a number of investigation [5]. [7] proposed a formalised process model for the conducting of computer forensics field triage – pushing some of the traditional forensic laboratory investigative steps to the field. [2] proposed a tiered methodology to enable based digital forensic triage capabilities by non-digital evidence specialists in order to facilitate the betting spending of expert time.

- Remote Acquisition – Many national police forces have a limited group of trained digital forensic evidence handling specialists [2]. Combining the deployment of digital forensic first responders with a remote evidence acquisition tool, such as that described by [9], would facilitate a more efficient utilisation of trained forensic experts. Uploading disk images from the field would facilitate forensic analysts to stay in the laboratory processing case work as opposed to wasting time out in the field.

- Automation – Expert experience and knowledge is predominantly used in conjunction with standard, and often basic, digital forensic tools to conduct the majority of digital investigations. While automation is an obvious improvement over the typical approach, it is not without its own challenges. [3] outline a number a challenges with automation in the digital forensic process including, with many centring around automated systems having a lack of knowledge capability of case specific information.

B. Data Deduplication

A deduplicated evidence storage system, such as those described by Watkins et al. [12] and Scanlon [8], can facilitate expedited evidence acquisition. Each unique file encountered need only be stored, indexed, analysed, and annotated once on the system. Eliminating non-pertinent, benign files during the acquisition phase of the investigation would immensely reduce the acquisition time (e.g., operating system, application, previously acquired non-incriminating files, etc.). This could greatly expedite pertinent information being available to the detectives working on the case as early as possible in the investigation. In order for any evidence to be court admissible, a forensically sound entire disk image would need to be reconstructible from the deduplicated data store, improving upon the system proposed by Watkins et al. [12]. Employing such a system would also facilitate a cloud-to-cloud based storage event monitoring of virtual systems as merely the changes of the virtual storage would need to be stored between each acquisition.

C. Digital Forensics as a Service (DFaaS)

Digital Forensics as a Service (DFaaS) is a cloud-based, modernised extension of current digital forensic investigative methods. While still very much in its infancy, some progress has been made in its development. In 2010, the Netherlands Forensic Institute (NFI) implemented a DFaaS solution in order to combat the volume of backlogged cases in the country [11]. [11] describe the advantages of the current system including efficient resource management, enabling detectives to directly query the data, improving the turnaround time between forming a hypothesis in an investigation its confirmation based on the evidence, and facilitating easier collaboration between detectives working on the same case through annotation and shared knowledge.

While the aforementioned DFaaS system is a significant step in the right direction, many improvements to the current model could greatly expedite and improve upon the current process. This includes improving the functionality available to the case detectives, improving its current indexing capabilities and on-the-fly identification of incriminating evidence during the acquisition process [11].

III. SYSTEM DESIGN

A. Acquisition

Figure 1 shows an overview of the disk image acquisition phase of the proposed system. The acquisition phase follows the traditional model initially requiring the crime scene processing and seizure of pertinent equipment. Once the identification and seizure phases are complete, there are two usage scenarios for employing the proposed data deduplication system. The first scenario involves the traditional process: bring the devices back to a forensic laboratory and imaging the storage media from there. Most forensic laboratories would use a network attached storage system (NAS) for storing complete disk images. In this LAN-based scenario, the data would instead be directed to the deduplicated system. The second scenario is a new paradigm made more feasible through the use of a deduplicated forensic model: a remote evidence acquisition from the field or local forensic laboratory back to a central evidence storage and processing system, such as a national law enforcement DFaaS system. In this scenario, the digital evidence is gathered and transmitted over the Internet in a secure and verifiable manner.

Irrespective of which usage scenario is necessary, the actual acquisition process is similar. Once the devices are seized and ready for imaging, they are connected to a forensic workstation using a write-blocker. The client side application initially
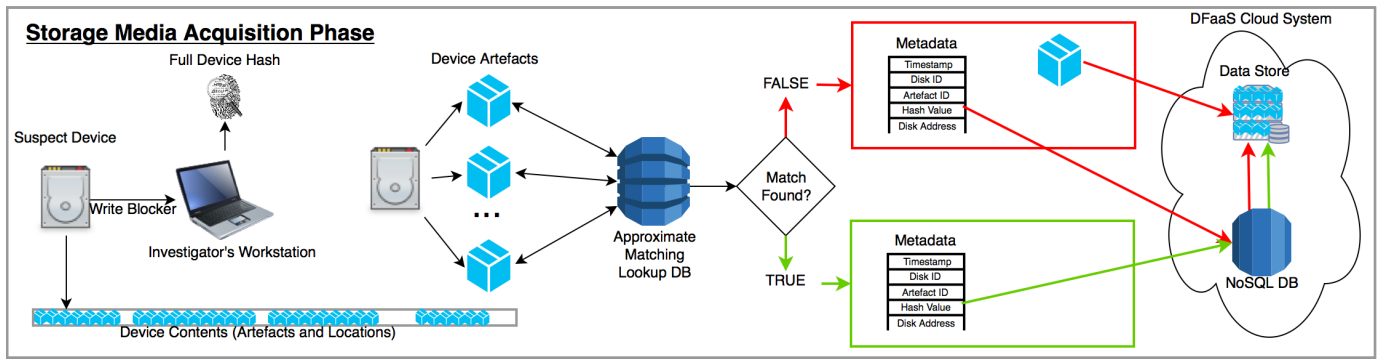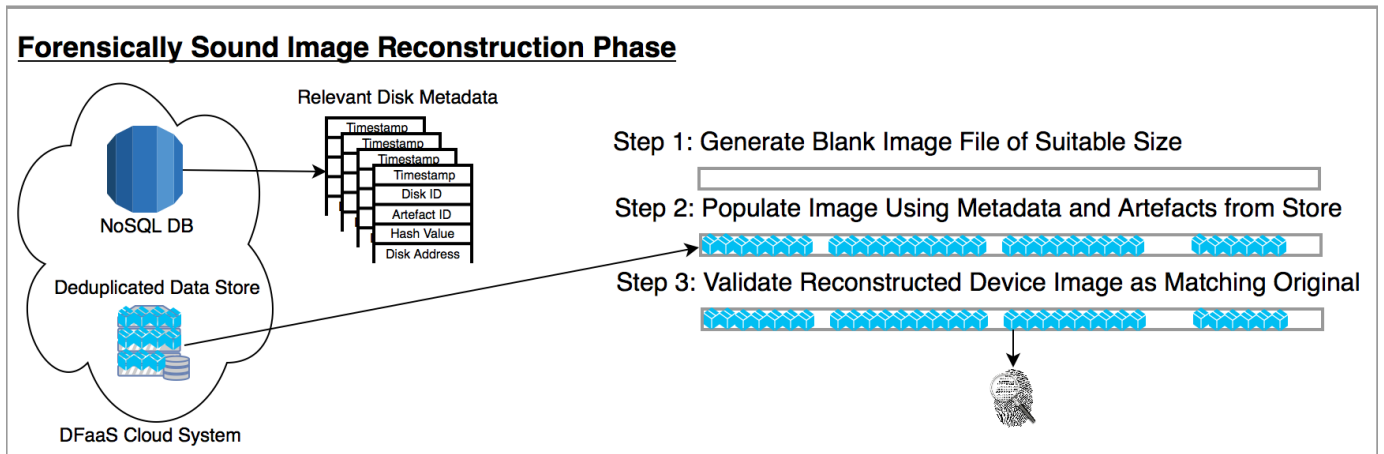
Fig. 1: Disk Image Acquisition Phase

Fig. 2: Complete Disk Image Reconstruction Phase

requires the identification and hashing of all storage device artefacts (partitions, files, slackspace, etc.) and compares each of these with the database of all previously acquired files. This comparison process can also be greatly expedited through the deployment of an approximate matching database. This approximate matching database can be frequently updated and generated on the server-side as files are added to the system. If a match is found, the associated metadata for this acquisition is recorded in the central database including the filename, creation and modification dates, the disk address, its fragmentation, and associated acquisition data such as the disk ID, case ID, investigator ID, etc. This metadata is linked to the deduplicated artefact stored on the server-side. If no such match is found, the same metadata is recorded alongside a transmission of a copy of the artefact itself to the server-side file repository.

In practice, local hashing, local metadata collection, comparison against the central repository, file uploading and indexing can all happen in a largely multi-threaded simultaneous fashion. Each file that is found to already exist on the server does not need to be reacquired effectively eliminating the transmission of this file (over the LAN or Internet) to the repository. However, in this model, the bottleneck of the disk-

read speed is still present as each artefact must be hashed in order to determine its duplicate status. Fuzzy hashing techniques could be integrated into the system, which would greatly improve the hashing speed, while sacrificing a small amount of accuracy. As such, the overall throughput during the acquisition speed of the data from the disk can be greater than the read speed of the disk.

B. Disk Image Reconstruction

When it comes to the reconstruction of an entire disk image from the deduplicated data storage, it is a matter of gathering all of the metadata associated with a specific acquisition and populating the disk image file artefact by artefact, as can be seen in Figure 2. A full-size blank disk image is first created as a staging area. Subsequently, each artefact is retrieved from the deduplicated data store and added into the disk image at the corresponding offset. Using this process, a complete disk image can be reconstructed with all of the corresponding artefacts found during the acquisition step. This reconstructed image would be verifiable as a true copy of the original through entire disk image hashing, e.g., to verify the court admissibility of the reconstructed evidence.
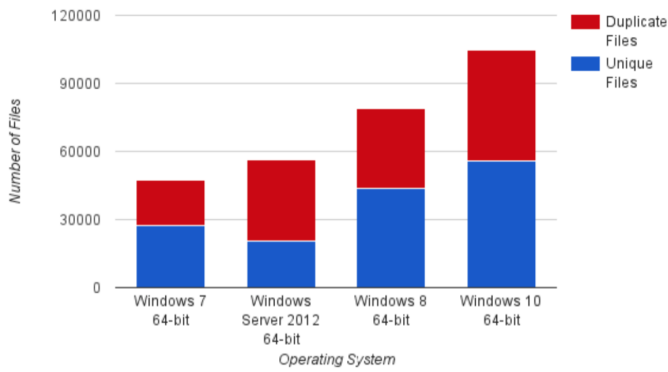
Fig. 3: Order of acquisitions to the deduplicated forensic storage system from L-R

Fig. 4: Comparison of Initial Acquisition vs. Second Acquisition Time per bare Operating System Install

IV. PRELIMINARY EXPERIMENTATION AND RESULTS

A. Experimental Setup

For the purposes of proving the viability of the proposed data deduplication system, a prototype client-server system was developed in Python using `pytsk` (python bindings for The Sleuth Kit). This section outlines some preliminary results of the acquisition phase using the system. `pytsk` facilities the processing of numerous file systems in python including NTFS, FAT12, FAT16, FAT32, exFAT, UFS1, UFS1b, UFS2, Ext2, Ext3, Ext4, SWAP, RAW, ISO9660, HFS and YAFFS2, but for the purposes of the experimentation presented in this paper, NTFS was used.

Four base virtual machines were created with separate virtual hard disks in place for each of the 64-bit operating systems Windows 7, Windows Server 2012, Windows 8 and Windows 10. The experiments below were conducted over the Internet (simulating the aforementioned remote acquisition scenario).

B. Initial Image Acquisition

Figure 3 shows the order of acquisition of the base images from left to right into a bare prototype, i.e., at the start of the Windows 7 acquisition, there was no files in the system. The level of duplication is portrayed in red, with the number of files identified during the Windows 7 duplication being those contained on the single disk, e.g., one file, `prnca00z.inf_loc`, was found to occur 89 times on the system. Subsequently, the base disk images relating to the other operating systems were added to the system. The runtime for these initial operating system acquisitions are highlighted in Table 1. As might be expected, the overall runtime of these acquisitions is relatively slow compared to what one might anticipate from the traditional alternative due to the hashing and comparison steps and a lack of substantial artefact elimination.

Figure 4 shows the improvement of a second acquisition of these base images with 100% duplication. The runtime is greatly improved, however there is significant room for further improvement as the processing and evidence elimination
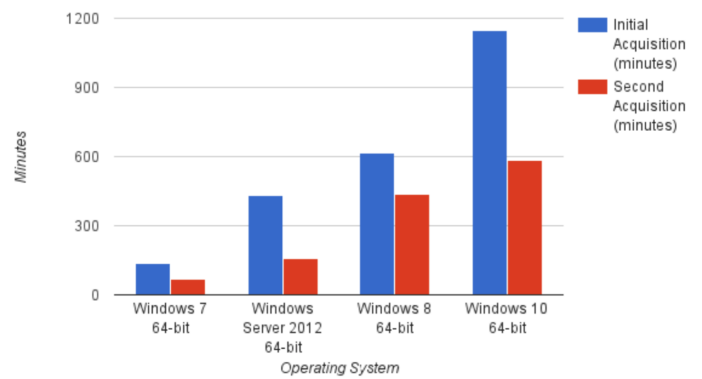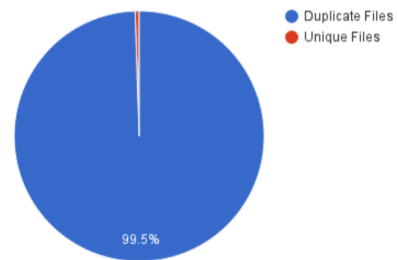

Fig. 5: Measurement of Modified Files between Boot Cycles of a Windows 10 Install

phases used for testing were running single threaded. The precise time measurements for this test are also outlined in Table 1.

C. Detection of Minor Modifications

For the purposes of this experiment, the Windows 10 system was booted up and used for regular Internet usage for a period of 30 minutes. In that time, a total of 0.5% of the files (550) on the system were modified and identified when the deduplication system reacquired the data. For reference, this acquisition was added to a database already containing Windows 10 acquisitions. As a result, 99.5% of the files on the disk were found to be duplicates in the system and hence, not reacquired.

V. CONCLUDING REMARKS

While the initial acquisition to a bare system takes longer than what would be expected from the traditional alternative (as would be expected for each new operating system added to any deduplicated storage, i.e., maximum artefact uniqueness), subsequent acquisitions of the same machine or additional machines sharing many common files are greatly improved. As more and more acquisitions are fed into the system, the acquisition and analysis time required for each new device analysed would correspondingly decrease, i.e., higher duplicate rates would be achieved. Such a technique could be used for the monitoring of virtual machines or cloud

Operating System	Initial Acquisition (Minutes)	Second Acquisition (Minutes)
Windows 7 64-bit	139.14	66.87
Windows Server 2012 64-bit	429.42	158.35
Windows 8 64-bit	615.29	439.44
Windows 10 64-bit	1148.35	583.69

TABLE I: Initial acquisition time for each disk image compared against its reacquisition

instances whereby the resolution of the captures could be much higher than any of the alternatives available. More intelligent deduplicated evidence data storage and analysis techniques can help eliminate duplicated processing and duplicated expert analysis of previously content. Once a file has been analysed and determined to be incriminating, an alert can be generated during the acquisition phase to notify the investigator that illegal content has been discovered at the earliest stage possible in the investigation.

A. Future Work

The system outlined as part of this paper is in its infancy and there is much work to be done to improve the system and increase its functionality, performance and viability as a true replacement for the traditional digital forensic approach. A number of these avenues for research and development are outlined below:

- Improvements to the processing capabilities of the prototype system in terms of multi-threading, fuzzy hashing, automated evidence classification, etc.
- Combining expert contributed analysis of acquired files facilitating automated blacklisting or whitelisting of suspect devices. Functionality such as this can greatly help towards the elimination of duplicated efforts in the analysis of content.
- Using the proposed deduplication system can also greatly expedite the acquisition of digital evidence from hash-based file-synchronisation services, such as BitTorrent Sync or Syncthing [10, 6]. Given that these tools rely on frequent hashing to determine when a synchronisation is required, those same hashes can be used for artefact acquisition and elimination purposes.

REFERENCES

[1] Eoghan Casey, Monique Ferraro, and Lam Nguyen. Investigation delayed is justice denied: proposals for expediting forensic examinations of digital evidence. *Journal of forensic sciences*, 54(6):1353–1364, 2009.

[2] Ben Hitchcock, Nhien-An Le-Khac, and Mark Scanlon. Tiered forensic methodology model for digital field triage by non-digital evidence specialists. *Digital Investigation*, 16(S1):75–85, 03 2016. ISSN 1742-2876. Proceedings of the Third Annual DFRWS Europe.

[3] Joshua I James and Pavel Gladyshev. Challenges with automation in digital forensic investigations. *arXiv preprint arXiv:1303.4498*, 2013.

[4] David Lillis, Brett Becker, Tadhg O'Sullivan, and Mark Scanlon. Current Challenges and Future Research Areas for Digital Forensic Investigation. 05 2016.

[5] Philip Penrose, William J Buchanan, and Richard Macfarlane. Fast contraband detection in large capacity disk drives. *Digital Investigation*, 12:S22–S29, 2015.

[6] Conor Quinn, Mark Scanlon, Jason Farina, and M-Tahar Kechadi. Forensic analysis and remote evidence recovery from syncthing: An open source decentralised file synchronisation utility. In *Digital Forensics and Cyber Crime*, pages 85–99. Springer, 2015.

[7] Marcus K Rogers, James Goldman, Rick Mislan, Timothy Wedge, and Steve Debrota. Computer forensics field triage process model. In *Proceedings of the conference on Digital Forensics, Security and Law*, page 27. Association of Digital Forensics, Security and Law, 2006.

[8] Mark Scanlon. Battling the Digital Forensic Backlog through Data Deduplication. In *Proceedings of the 6th IEEE International Conference on Innovative Computing Technologies (INTECH 2016)*, Dublin, Ireland, 08 2016. IEEE.

[9] Mark Scanlon and M-Tahar Kechadi. Online acquisition of digital forensic evidence. In *Proceedings of International Conference on Digital Forensics and Cyber Crime (ICDF2C 2009)*, pages 122–131. Springer, Albany, New York, USA, September 2009.

[10] Mark Scanlon, Jason Farina, and M-Tahar Kechadi. Network investigation methodology for bittorrent sync: A peer-to-peer based file synchronisation service. *Computers and Security*, 54:27 – 43, 10 2015. ISSN 0167-4048. doi: http://dx.doi.org/10.1016/j.cose.2015.05.003. URL http://www.sciencedirect.com/science/article/pii/S016740481500067X.

[11] RB Van Baar, HMA van Beek, and EJ van Eijk. Digital forensics as a service: A game changer. *Digital Investigation*, 11:S54–S62, 2014.

[12] Kathryn Watkins, Mike McWhorte, Jeff Long, and Bill Hill. Teleporter: An Analytically and Forensically Sound Duplicate Transfer System. *Digital Investigation*, 6:S43–S47, 2009.