NEW APPROACHES TO DIGITAL EVIDENCE ACQUISITION AND ANALYSIS

BY **MARTIN NOVAK, JONATHAN GRIER,** AND **DANIEL GONZALES**

Two NIJ-supported projects offer innovative ways to process digital evidence.

Computers are used to commit crime, but with the burgeoning science of digital evidence forensics, law enforcement can now use computers to fight crime.

Digital evidence is information stored or transmitted in binary form that may be relied on in court. It can be found on a computer hard drive, a mobile phone, a CD, and a flash card in a digital camera, among other places. Digital evidence is commonly associated with electronic crime, or e-crime, such as child pornography or credit card fraud. However, digital evidence is now used to prosecute all types of crimes, not just e-crime. For example, suspects' email or mobile phone files might contain critical evidence regarding their intent, their whereabouts at the time of a crime, and their relationship with other suspects.

In an effort to fight e-crime and to collect relevant digital evidence for all crimes, law enforcement agencies are incorporating the collection and analysis of digital evidence into their infrastructure.

Digital forensics essentially involves a three-step, sequential process:[1]

1. Seizing the media.
2. Acquiring the media; that is, creating a forensic image of the media for examination.
3. Analyzing the forensic image of the original media. This ensures that the original media are not modified during analysis and helps preserve the probative value of the evidence.

Sifting Collectors has the potential to significantly reduce digital forensics backlogs and quickly get valuable evidence to the people who need it.

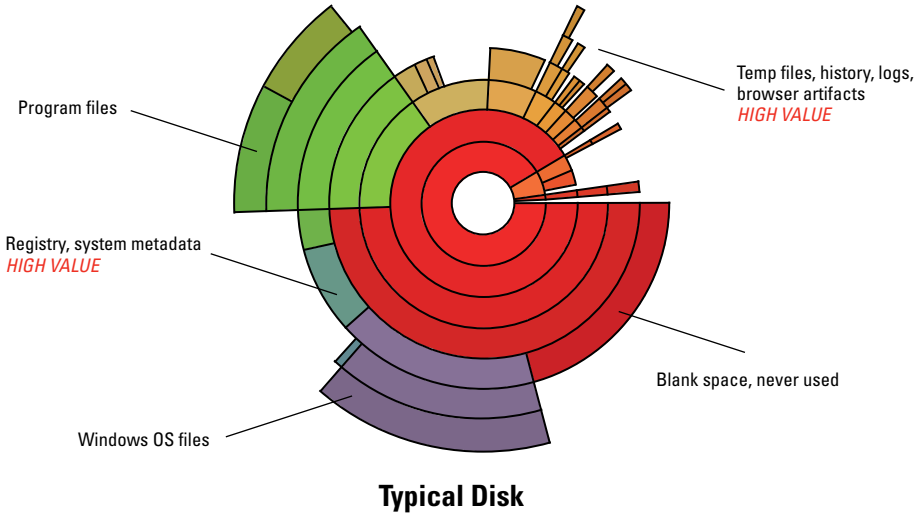
Large-capacity media typically seized as evidence in a criminal investigation, such as computer hard drives and external drives, may be 1 terabyte (TB) or larger. This is equivalent to about 17,000 hours of compressed recorded audio. Today, media can be acquired forensically at approximately 1.5 gigabytes (GB) per minute. The forensically acquired media are stored in a RAW image format, which results in a bit-for-bit copy of the data contained in the original media without any additions or deletions, even for the portions of the media that do not contain data. This means that a 1 TB hard drive will take approximately 11 hours for forensic acquisition.[2] Although this method captures all possible data stored in a piece of digital media, it is time-consuming and creates backlogs. In 2014, there were 7,800 backlogged cases involving digital forensics in publicly funded forensic crime labs.[3]

To help address these challenges, NIJ funded two projects in 2014: Grier Forensics received an award to develop a new approach to acquiring digital media, and RAND Corporation received an award to work on an innovative means for analyzing digital media. Four years later, these software applications are coming to fruition.

Identifying Disk Regions That May Contain Evidence

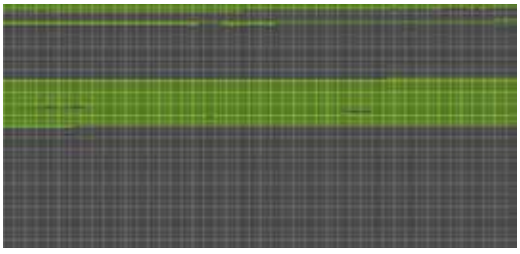
Traditional disk acquisition tools produce a disk image that is a bit-for-bit duplicate of the original media. Therefore, if a piece of acquired media is 2 TB in size, then the disk image produced will also be 2 TB in size. The disk image will include all regions of the original media, even those that are blank, unused, or irrelevant to the investigation. It will also include large portions devoted to operating systems (e.g., Windows 10 or Mac OSX), third-party applications, and programs supplied by vendors such as Microsoft or Apple (see exhibit 1).

Exhibit 1. Typical Disk Regions



Source: Courtesy of Grier Forensics.

Exhibit 2. Visualization of Disk Regions



Source: Courtesy of Grier Forensics.

For some cases, such as software piracy, it is important to collect these programs so investigators can understand the computer's original environment. However, for the vast majority of cases, these regions are not important. For most computer forensic investigations, the evidence lies in the user's documents, emails, internet history, and any downloaded illicit images.

Grier Forensics proposed a novel approach that images only those regions of a disk that may contain evidence. Called the Rapid Forensic Acquisition of Large Media with Sifting Collectors (Sifting Collectors for short), this software application bypasses regions that contain exclusively third-party, unmodified applications and, instead, zeroes in on the regions that contain data, artifacts, and other evidence. (The software can be easily configured to collect third-party applications when necessary for certain types of cases.)

Exhibit 2 is a visualization of disk regions generated by the Sifting Collectors diagnostic package. The green areas represent user-created files and the black areas represent portions of the media that have never been used.

Sifting Collectors has the potential to significantly reduce digital forensics backlogs and quickly get valuable evidence to the people who need it. In laboratory testing,[4] it accelerated the imaging process by three to 13 times while still yielding 95 to 100 percent of the evidence.

Sifting Collectors is designed to drop right into existing practices. The software creates an industry-standard forensic file — known as an "E01 file" — that is accessible from standard forensic tools, just like current imaging methods.[5] Grier Forensics is working with major forensics suite manufacturers to allow Sifting Collectors to work seamlessly with their existing tools.

Potential Limitations of Sifting Collectors

Perhaps the most significant drawback of Sifting Collectors is that, unlike traditional imaging, it does not collect the entire disk. Instead, Sifting Collectors discovers which regions of the disk may contain evidence and which do not.

This might not be a significant drawback, however. Digital evidence is typically handled in one of two ways:

- The investigators seize and maintain the original evidence (i.e., the disk). This is the typical practice of law enforcement organizations.
- The original evidence is not seized, and access to collect evidence is available only for a limited duration. This is common in cases involving ongoing intelligence gathering — for example, when law enforcement has a valid search warrant to collect evidence but, because of an ongoing investigation, does not plan to seize the evidence.

In the second scenario, computer forensics examiners have a limited time window for entering the site and collecting as much evidence as possible. Consequently, they will focus only on the most valuable devices and then image each device, spending more than half of their time collecting unmodified regions (as described above). Sifting Collectors would allow them to accelerate the process and collect evidence from many more devices. Either way, given the limited time window, it is difficult to collect all digital evidence. The choice for the computer forensics examiner is whether to collect

all regions, including blanks, from a small number of devices or to collect only modified regions containing evidence from a large number of devices. Sifting Collectors allows examiners to make that choice.

When investigators retain the original evidence, the mitigation is even simpler: Sifting Collectors allows users to collect and analyze disk regions expected to contain evidence. It allows them to acquire evidence quickly and start the case more rapidly, and it potentially reduces case backlogs. If, at any time, users need to analyze other regions, they can go back to the original and collect those regions.

Another potential drawback concerns hash verification — using an electronic signature or verification code, known as a hash, to verify that a disk image matches the original evidence disk. Existing methods of hash verification depend on verifying the entire disk and thus are not compatible with Sifting Collectors. However, this problem is not limited to Sifting Collectors; modern, solid-state drives (SSDs) are often incompatible with hash verification because certain SSD regions are unstable due to maintenance operations. In both cases, the solution is the same: moving from disk-based verification to more granular verification strategies. As the industry adopts newer verification strategies to accommodate SSDs, Sifting Collectors will likely benefit as well.

The process that Sifting Collectors uses to analyze the disk and distinguish relevant regions from unmodified or irrelevant ones takes time. The amount of time varies greatly based on the disk, but it could be up to 10 percent of the imaging time. This means that if Sifting Collectors determines that it is necessary to collect the entire disk or nearly all of it, the software will not save the user any time and will, in fact, be somewhat slower than current imaging methods. To help mitigate this, Grier Forensics is using advanced parallel processing, concurrency, and compression algorithms. However, even with these modifications, Sifting Collectors will end up being slightly slower than traditional imaging in cases where nearly all of the disk is collected.

Perhaps the drawback that is likely to cause the most resistance is simply that Sifting Collectors necessitates a break with current practice. Indeed, reluctance to change current practice will be a substantial obstacle to overcome if Sifting Collectors is to achieve widespread adoption.

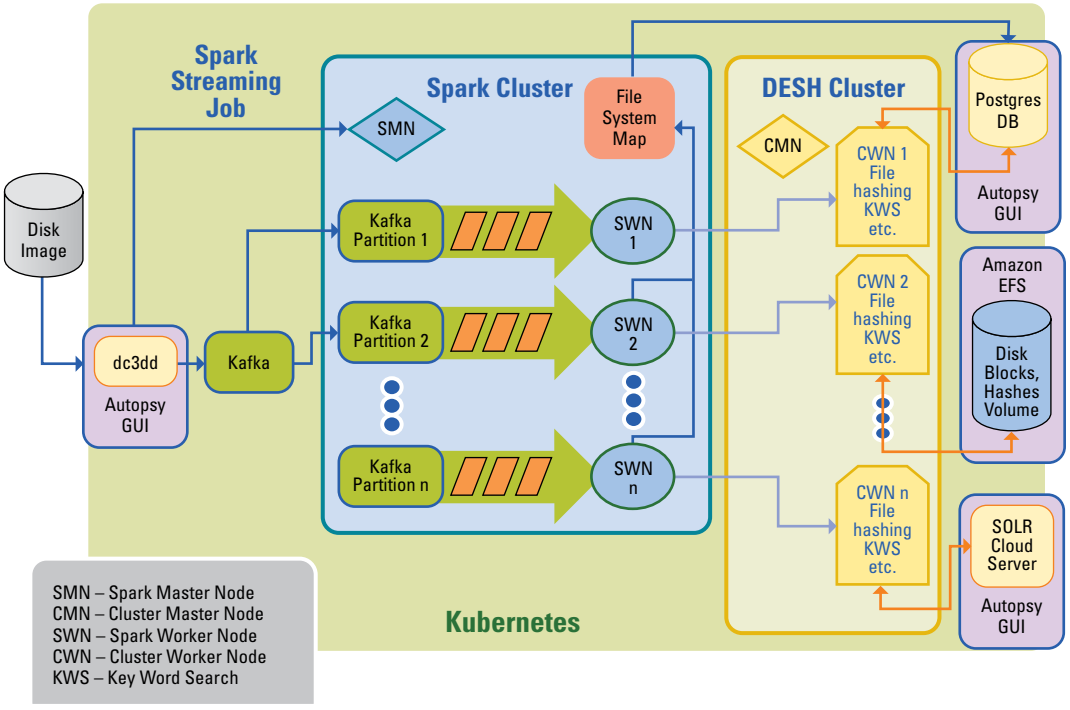
Accelerating Digital Forensics Analysis

Each year, the time it takes to conduct digital forensics investigations increases as the size of hard drives continues to increase. With NIJ support, RAND has developed an open-source digital forensics processing application designed to reduce the time required to conduct forensically sound investigations of data stored on desktop computers. The application, called the Digital Forensics Compute Cluster (DFORC2), takes advantage of the parallel-processing capability of stand-alone high-performance servers or cloud-computing environments (e.g., it has been tested on the Amazon Web Services cloud).

DFORC2 is an open-source project. It uses open-source software packages such as dc3dd,[6] Apache Kafka,[7] and Apache Spark.[8] Users interact with DFORC2 through Autopsy, an open-source digital forensics tool that is widely used by law enforcement and other government agencies and is designed to hide complexity from the user. RAND has designed DFORC2 so the application can also use the Kubernetes Cluster Manager,[9] an open-source project that provides auto-scaling capabilities when deployed to appropriate cloud-computing services. (See exhibit 3 for a detailed description of how DFORC2 works.)

The primary advantage of DFORC2 is that it will significantly reduce the time required to ingest and process digital evidence. DFORC2's speed advantage, however, will depend on two factors. The first factor is the speed and memory of the server. For smaller servers (those with 16 GB of RAM or less and an older microprocessor), the original stand-alone version of Autopsy will perform better than DFORC2. On a larger server (one with 28 GB of RAM or more and a new high-end multicore microprocessor), DFORC2 will be faster.

Exhibit 3. DFORC2 System Architecture



Source: Courtesy of RAND Corporation.

Note: A compute cluster has its resources organized into a cluster manager and worker nodes. Worker nodes perform computing tasks assigned to them by the cluster manager. DFORC2 ingests data from the hard drive (using dc3dd) and streams it in "blocks" to the Apache Spark cluster. Apache Spark worker nodes search for logical file metadata and send their findings to the PostgreSQL database. Data blocks are hashed before and after receipt to ensure integrity. As the streamed data are received, worker nodes in a second cluster, the Digital Evidence Search and Hash (DESH) cluster, identify and reconstruct "complete" files and process these files using local copies of the Autopsy application. An essential part of the core workflow is the reconstruction of the master file system during the file ingestion process. This is done by the Apache Spark cluster, during rather than after file ingestion, to speed up the forensics analysis process. The master file system map or table and logical file metadata are stored in the PostgreSQL database.

The second factor is the number of worker nodes that can be allocated to the clusters. DFORC2 organizes resources into a cluster manager and worker nodes. Worker nodes perform computing tasks assigned to them by the cluster manager. More worker nodes will significantly reduce evidence ingest and processing times. However, there is a limit to the number of worker nodes that can be implemented on a server, even one that is equipped with a state-of-the-art multicore microprocessor. To get the full benefit of large numbers of worker nodes, the cloud-based version of DFORC2 is needed; the Kubernetes Cluster Manager can spread data-processing tasks over multiple machines in the cloud.

Potential Limitations of DFORC2

The first potential limitation is the complexity of the current prototype. Currently, distributed computing expertise is needed to set up and implement the stand-alone version of DFORC2. RAND is working to simplify its installation on a stand-alone server.

A different set of complex tasks is required to implement DFORC2 in a commercial cloud. Although the Kubernetes Cluster Manager simplifies much of the system's internal setup and configuration, a number of complex steps are required to ensure

secure communications with a DFORC2 cloud installation.

In developing its prototype, RAND is using the Amazon Web Services computing cloud. It communicates with the DFORC2 prototype through the firewalls protecting RAND's enterprise network. RAND has had to work through a number of security and firewall exception issues to enable the smooth installation and startup of DFORC2 in Amazon Web Services. This is another setup and installation issue that RAND is working to simplify so law enforcement agencies can securely access their own DFORC2 cloud installations from their enterprise networks.

Another potential concern with the use of DFORC2 in criminal investigations is the chain of custody for evidence when commercial cloud-computing services are used to process and store evidence. Additional processing and communication steps are involved when using DFORC2.[10] RAND is conducting a chain-of-custody analysis to strengthen the integrity of the digital forensics processing paths used by DFORC2 in a commercial cloud. Additional cloud security features can also be enabled to protect user data and strengthen the chain of custody in the cloud.

Finally, an additional source of concern is how compute clusters handle data. The chain-of-custody analysis now underway will examine this issue and will include a comprehensive review of the distributed computing software components used in DFORC2.

Need for Evaluation

With the support of NIJ, Grier Forensics and RAND are moving the field forward by developing new means for processing digital evidence. Grier Forensics' Sifting Collectors provides the next step in the evolution of evidence acquisition. RAND's DFORC2 combines the power of compute clusters with open-source forensic analysis software to process evidence more efficiently.

Both of these projects introduce new paradigms for the acquisition and analysis of digital evidence. Whether the criminal justice community accepts these approaches will depend on the admissibility of the evidence each produces. That admissibility will ultimately be determined by the threshold tests of the *Daubert* standard in court. These new approaches will need to be independently tested, validated, and subjected to peer review. Known error rates and the standards and protocols for the execution of their methodologies will need to be determined. In addition, the relevant scientific community must accept them.

RAND will release DFORC2 software code to their law enforcement partners and members of the digital forensics research community in the near future. They will test it, find bugs, and improve the code. Eventually, it will be released as an open-source project.

Grier Forensics will release Sifting Collectors to their law enforcement partners for field trials to verify its preliminary laboratory findings with real cases. It recently benchmarked Sifting Collectors against conventional forensic imaging technology and found that Sifting Collectors was two to 14 times as fast as conventional imaging technology, depending on the mode and the source disk, and produced an image file requiring one-third the storage space — and it still achieved 99.73 percent comprehensiveness (as measured by a third-party tool).

Meanwhile, NIJ plans to have both DFORC2 and Sifting Collectors independently tested by the NIJ-supported National Criminal Justice Technology Research, Test and Evaluation Center, which is hosted by the Applied Physics Laboratory at Johns Hopkins University.

About the Authors

Martin Novak is a senior computer scientist in NIJ's Office of Science and Technology. **Jonathan Grier** has performed security research, consulting, and investigation for more than 15 years. He developed new security technology for the Defense Advanced Research Projects Agency, the Massachusetts Institute

of Technology Lincoln Laboratory, the National Institute of Standards and Technology, and the United States Air Force. **Daniel Gonzales,** Ph.D., is a senior physical scientist at RAND Corporation. He has expertise in command, control, and communications systems; electronic warfare; cybersecurity; digital forensics; critical infrastructure protection; and emergency communications.

For More Information

Read the results of an NIJ-sponsored research effort to identify and prioritize criminal justice needs related to digital evidence collection, management, analysis, and use at NIJ.ojp.gov, keyword: 248770.

Read the findings of an NIJ-sponsored expert panel on the challenges facing law enforcement when accessing data in remote data centers at https://www.rand.org/pubs/research_reports/RR2240.html.

This article discusses the following grants:

- "Rapid Forensic Acquisition of Large Media with Sifting Collectors," grant number 2014-IJ-CX-K001
- "Rapid Forensic Acquisition of Large Media with Sifting Collectors," grant number 2014-IJ-CX-K401
- "Accelerating Digital Evidence Analysis Using Recent Advances In Parallel Processing," grant number 2014-IJ-CX-K102

Notes

1. National Institute of Justice funding opportunity, "New Approaches to Digital Evidence Processing and Storage," Grants.gov announcement number NIJ-2014-3727, posted February 6, 2014, https://www.ncjrs.gov/pdffiles1/nij/sl001078.pdf.
2. Steven Branigan, "Identifying and Removing Bottlenecks in Computer Forensic Imaging," poster session presented at NIJ Advanced Technology Conference, Washington, DC, June 2012.
3. Matthew R. Durose, Andrea M. Burch, Kelly Walsh, and Emily Tiry, *Publicly Funded Forensic Crime Laboratories: Resources and Services, 2014* (Washington, DC: U.S. Department of Justice, Bureau of Justice Statistics, November 2016), NCJ 250151, https://www.bjs.gov/content/pub/pdf/pffclrs14.pdf.
4. The tests used disk images from DigitalCorpora.org, a website of digital corpora for use in computer forensics education research that is funded through the National Science Foundation.
5. Simson L. Garfinkel, David J. Malan, Karl-Alexander Dubec, Christopher C. Stevens, and Cecile Pham, "Advanced Forensic Format: An Open Extensible Format for Disk Imaging," in *Advances in Digital Forensics II,* ed. Martin S. Olivier and Sujeet Shenoi (New York: Springer, 2006), 13-27.
6. The application dc3dd, created by the Department of Defense's Cyber Crime Center, is capable of hashing files and disk blocks "on the fly" as a disk is being read. The application can be downloaded at SourceForge.
7. Apache Kafka is an open-source stream processing platform that provides a unified, high-throughput, low-latency platform for handling real-time data feeds.
8. Apache Spark provides an interface for programming entire clusters with implicit data parallelism and fault tolerance.
9. Kubernetes Cluster Manager is an open-source platform that automates deployment, scaling, and operations of applications on compute clusters. If the Kubernetes Cluster Manager is not used (e.g., if DFORC2 is deployed to a single server), then the user will fix the number of worker nodes performing forensics analysis tasks at runtime. Because of this, digital forensics analysts using DFORC2 would have to estimate the number of Apache Spark and Digital Evidence Search and Hash cluster worker nodes needed for a specific size of hard disk and for a specific type of investigation. The number of compute nodes needed could depend on many factors, which the analyst may not know before the investigation is started. This limitation would likely require the analyst to overprovision the cloud compute cluster to ensure timely processing of the evidence. The Kubernetes Cluster Manager solves this problem. It is designed to deploy or shut down cluster computing resources, depending on the level of demand on each virtual machine. Furthermore, it is compatible with a wide range of cloud-computing environments. The Kubernetes Cluster Manager can deploy applications on demand, scale applications while processes are running in containers (i.e., add additional worker nodes to compute tasks), and optimize hardware resources and limit costs by using only the resources needed.

10. The DFORC2 chain of custody relies on cryptographic hashes to verify the content of disk blocks and logical files found on the hard disk that is the subject of investigation. All disk blocks are hashed twice, first by dc3dd when the disk is read into DFORC2. This hashing takes place outside the cloud, on a local computer that is used to ingest the hard disk and stream it into the cloud. Autopsy then hashes the disk blocks a second time inside the cloud. These two hashes can be compared to prove that the copy of the disk in the cloud is identical to the disk block ingested from the original piece of evidence. Logical files are not hashed during data ingestion. However, they can be hashed on the local computer using an accepted standard digital forensics tool if this is required to verify evidence found in a specific file by DFORC2 in the cloud. All logical file hashes are retained by DFORC2 in the cloud to enable the analyst to trace the chain of custody for specific pieces of evidence on an as-needed basis.

Image source: PeterPhoto123/Shutterstock

NCJ 250700

Cite this article as: Martin Novak, Jonathan Grier, and Daniel Gonzalez, "New Approaches to Digital Evidence Acquisition and Analysis," *NIJ Journal* 280, January 2019, https://www.nij.gov/journals/280/pages/new-approaches-to-digital-evidence-acquisition-and-analysis.aspx.