

By GOLDEN G. RICHARD III *and* VASSIL ROUSSEV

Next-Generation DIGITAL FORENSICS

The digital forensics community requires new tools and strategies for the rapid turnaround of large forensic targets.

Digital forensics investigators are experiencing an increase in both the number and complexity of cases that require their attention. The number of cases is rising for a variety of reasons, including better awareness of the capabilities of digital forensics techniques at all levels of law enforcement and in the private sector. The complexity of cases is growing because forensic targets with hundreds of gigabytes or terabytes of storage are becoming common and cases routinely involve more than a single computer.

Furthermore, longer queues of larger and more complicated cases don't relax the need for quick turnaround, especially in time-sensitive cases involving potential loss of life or property. This leaves digital forensics investigators in need of tools that are significantly better, both in richness of features and in speed. Providing practitioners with these tools will require a critical look at current offerings and significant levels of innovation on the part of researchers. This article discusses some shortcomings of the current generation of digital

Illustration *by* Chris Buzelli

forensics tools, previews some research addressing some of these shortcomings, and suggests areas that need additional work.

Most current digital forensics tools are designed to run on a single workstation, with the investigator issuing queries against copies of the acquired digital evidence. With current generation tools, the single workstation model works reasonably well and allows tolerable case turnaround times for small forensic targets (for example, < 40GB). For much larger targets, these tools are too slow to provide acceptable turnaround times.

One approach to handling larger targets is to carefully analyze available tools and consider substantial updates to their designs. Since many existing tools were originally released when much smaller targets were the norm, there is reason to believe this approach has merit; indeed, some limited yet compelling evidence confirms this belief [9]. Unlike much application software, where ever-faster hardware hides inefficient software design, wasted CPU cycles in a digital forensics tool literally mean that a critical case, such as one involving kidnapping or terrorist activity, remains unsolved for a longer period of time. Many tools may benefit from applying rules of thumb in the design of operating systems: avoiding unnecessary memory-to-memory copies and reduction of disk I/O, particularly write operations. This is the approach taken in [9] and experimental results show that substantial time savings are possible, at least in the target application (file carving, which is essentially retrieval of files based on analysis of file formats). Ultimately, however, even highly optimized tools, when run on single workstations, will be unable to process extremely large targets in a reasonable amount of time. Here, we discuss some other possible approaches for handling large targets and illuminate some key areas where additional research is needed.

BETTER ACQUISITION NEEDED

A typical digital forensics investigation consists of several major steps, including acquisition and duplication of digital evidence, analysis of evidence, and presentation of results. For large targets, all of the steps in an investigation become more burdensome, but acquisition and analysis are the most in need of improvement.

A major problem when investigating large targets is how to capture the essential data during acquisition,

when working copies of potential evidence sources are created. Smarter acquisition can reduce the amount of data that must be examined by targeting interesting evidence and leaving behind data that is known to have no evidentiary value. Data reduction of this sort can be based on a database of “known files,” including standard operating systems or application components. During acquisition, standard operating systems components and application files can be left behind, since in most cases they aren’t of interest.

Data reduction can also be based on file types, for example, retrieval of only image files in a child pornography case. Large-scale cryptographic hash databases for known files already exist—perhaps the most well-known example is the National Software Reference Library (NSRL), which contains hashes for over 31,000,000 files. Currently, such data reduction techniques are in common use only in the analysis phase of an investigation, once evidence acquisition has been completed.

As targets grow larger, it will be increasingly important to move data reduction facilities to the acquisition phase, to avoid lengthy evidence duplication times and to avoid overwhelming investigators with irrelevant data. There is a common belief that failing to duplicate and examine all available evidence may be open to legal challenge, since exculpatory evidence (in particular) may be missed. But in the future, complete bit-by-bit captures of huge targets may be completely impractical, in the same sense that capturing the state of an entire building is impractical in a (non-digital) forensics investigation involving a murder. Further, Kenneally et al. argue in [6] that careful extraction of only relevant evidence may actually reduce legal risk.

Another pressing issue for the evidence acquisition phase is better tools for live acquisition. An old debate exists in the digital forensics community about whether to immediately “pull the plug” on computers being seized for investigation. One side of the argument is that pulling the plug minimizes disturbance of stored evidence, which is mostly true. Unfortunately, in many cases, pulling the plug may destroy a significant amount of volatile evidence. In at least two different circumstances, the worst case scenario is complete loss or inaccessibility of *all* evidence on the target. In one case, loss of the RAM contents for a target may make any stored evidence on the target inaccessible, if encryption products are in use. Of

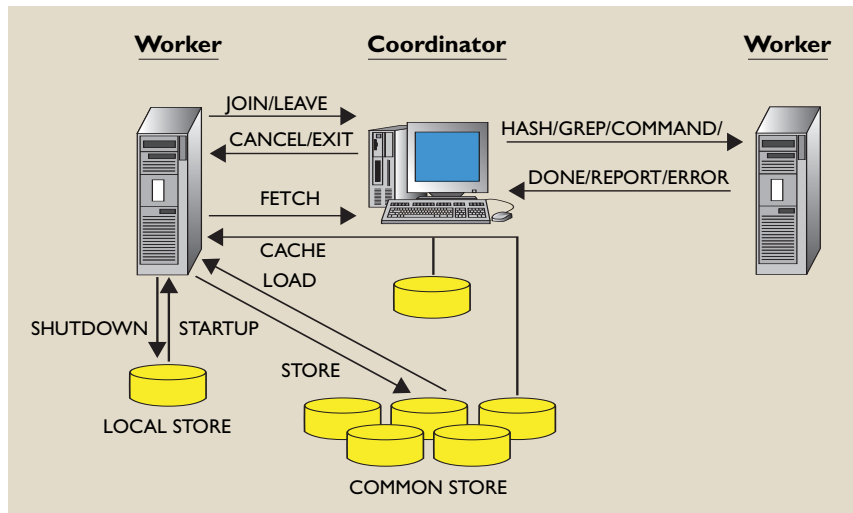
particular concern is “whole disk” encryption, where drive contents are decrypted on the fly. The decryption key, supplied by the user when the system boots and stored in RAM, may be recoverable only if the RAM contents is saved for later analysis. The second case occurs when a machine is running a completely volatile, bootable OS distribution, such as Knoppix. In this situation, a hard drive might not even be installed in the machine—all the interesting data is in RAM.

A difficulty in live acquisition is that if malicious software, such as a rootkit, has been installed on the target, it may be difficult to capture an accurate snapshot of the machine’s live state. Carrier [1] discusses a hardware-assisted solution for memory capture, but the hardware must have been installed before the last boot of the target. In many cases, such as investigating intrusions into corporate systems, this is not a major limitation, but the technique is not directly applicable to investigation of arbitrary systems. More research in this very important area is needed.

SPEEDING UP ANALYSIS: DISTRIBUTED COMPUTING

Most investigative techniques offered by today’s digital forensics tools—keyword searches, file classification, image thumbnailing—are I/O-bound. More sophisticated investigative techniques will be affected by both the limited I/O bandwidth of mechanical disks *and* requires substantial CPU resources. A distributed solution (for example, distributed digital forensics software executing on a Beowulf cluster), can address both I/O and processing constraints, using aggressive data caching techniques and performing investigative operations in parallel. Consider that with 2GB RAM per node, a 100-node cluster can easily cache a 200GB disk image completely in RAM. This supports loading a large disk image only once, distributing the data in the disk image among the nodes in the cluster, and then performing most operations against the cached data.

Since most digital forensics operations are file-centric, a natural way to distribute data in a computer cluster is on file boundaries. Very large files and large blocks of unallocated space (which frequently contain interesting evidence) are problematic and need to be split across nodes to make maximum use of the available RAM. Depending on the types of forensic oper-



Basic DELV (Distributed Environment for Large-scale Investigations) architecture. In a cluster, a single coordinator node performs group management functions, exports a user interface, and handles evidence distribution duties. A number of worker nodes cache digital evidence and accept commands from the coordinator to perform image thumbnailing, regular expression searches, steganography detection, and other operations in parallel. The architecture works on clusters with or without shared storage.

ations performed, it may be useful to ensure even distribution of files of certain types across the nodes. For example, thumbnailing operations can be executed more rapidly if more processors store image files, rather than having image files clustered on only a few nodes.

The prototype DELV (Distributed Environment for Large-scale Investigations) system, described in [10] and illustrated in the accompanying figure, uses these principles to achieve substantial speedup of forensic operations, even when using only a limited number of nodes. The availability of abundant CPU resources allows the creation of very responsive tools, allowing an investigator using DELV to issue many parallel queries and see results as they become available, rather than using a sequential query/response/query model, as is more common in current generation tools. To date, experiments with DELV show speedups well in excess of the concurrency factor for I/O bound operations and nearly linear speedups for CPU bound operations, such as steganography detection. More work is needed on an effective user interface for DELV, on fault tolerance facilities for handling node failures, and on a plug-in architecture for incorporating and parallelizing existing (sequential) tools.

BEYOND SIMPLE SPEEDUP: MORE SOPHISTICATED ANALYSIS

It’s not enough to perform smarter evidence acquisition and to use distributed computing to speed up the analysis phase of an investigation. Even if an

improved acquisition step substantially reduces the amount of data that must be processed, a huge amount of data must still be examined on a large target. And even if the speed of standard operations such as keyword searches, file classification, computation of cryptographic hashes, and generation of thumbnails is increased sufficiently to handle large targets, the investigative process remains too manual. For example, if the generation of image thumbnails is nearly instantaneous, reviewing the thumbnails to discover contraband is significantly time-consuming if the number of images is large. If a mechanism for automatically classifying and organizing the images was available, however, then an investigator could quickly eliminate image categories irrelevant to the case she is investigating.

Carrier [2] discusses automatic detection of evidentiary outliers by examining file attributes on a compromised honeypot. The goal is to find files that “stick out” because of their dissimilarity to other files in some set (for example, other files located in the same directory). Chin et al. [3] use content-based image recognition (CBIR) techniques to allow searching a forensic target for specific images that may have been subjected to a variety of image transformations, including resizing, rotation, and quality reduction. Unfortunately, use of similar techniques for content-based image clustering has so far failed to produce satisfactory results. Pal et al. [8] attempt to reconstruct complete images from a collection of fragments. This problem is NP-complete but the authors use several heuristic algorithms and the system works reasonably well for small collections of images. Finally, de Vel [4] and Novak et al. [7] present research on mining authorship information from text documents. In many cases, these research prototypes are deployed on uniprocessors and require a substantial amount of computation, so appropriate distributed algorithms must be developed once the basic techniques are refined.

CONCLUSION

The size of the targets that digital forensics investigators must process continues to grow and the current generation of digital forensics tools is already struggling to deal with even modest-sized targets. At a recent breakout session [5] on methods for dealing with large amounts of digital evidence, participants stated they are already trying to handle individual cases with between 200GB and 2TB of data. Virtually all of the participants expressed the need for dramatic improvements in the evidence acquisition and duplication process as well as more automated tools for analysis.

Ultimately, the digital forensics community not only requires better acquisition tools, but also better analysis tools that rely on distributed processing. To reduce case turnaround time, digital forensics investigators must be relieved of manual, time-consuming tasks, allowing them more time to think. Commodity compute clusters can not only dramatically reduce the waiting time for traditional forensic operations such as keyword searches, image thumbnailing, and file classification, but also provide an architecture for development of much more powerful analysis techniques. Such techniques include automated image analysis, multimedia processing, identification of evidentiary outliers, and correlation of sources of evidence. Tools to automate analysis of digital evidence are needed now—current tools are not prepared for the huge targets of tomorrow (or today). **C**

REFERENCES

1. Carrier, B. and Grand, J. A hardware-based memory acquisition procedure for digital investigations. *Digital Investigation* 1, 1 (Feb. 2004).
2. Carrier, B. and Spafford, E. Automated digital evidence target definition using outlier analysis and existing evidence. In *Proceedings of the 2005 Digital Forensics Research Workshop*.
3. Chin, Y., Roussev, V., Richard III, G.G. and Gao, Y. Content-based image retrieval for digital forensics. In *Proceedings of the International Conference on Digital Forensics (IFIP 2005)*.
4. de Vel, O.Y., Anderson, A., Corney, M. and Mohay, G.M. Mining email content for author identification forensics. *SIGMOD Record* 30, 4 (2001).
5. Digital Evidence Overload: Scalability and Automation. Breakout sessions at *The 5th Annual Digital Forensic Research Workshop* (Aug. 17–18, 2005, New Orleans, LA).
6. Kenneally, E. and Brown, C. Risk sensitive digital evidence collection. *Digital Investigation* 2, 2 (June 2005).
7. Novak, J., Raghavan, P. and Tomkins, A. Anti-aliasing on the Web. In *Proceedings of the 13th International Conference on the World Wide Web*, 2004.
8. Pal, A., Shanmugasundaram, K. and Memon, N. Automated reassembly of fragmented images. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2003.
9. Richard III, G.G. and Roussev, V. Scalpel: A frugal, high-performance file carver. In *Proceedings of the 2005 Digital Forensics Research Workshop*.
10. Roussev, V. and Richard III, G.G. Breaking the performance wall: The case for distributed digital forensics. In *Proceedings of the 2004 Digital Forensics Research Workshop*.

GOLDEN G. RICHARD III (golden@cs.uno.edu) is an associate professor in the Department of Computer Science at the University of New Orleans, where he teaches digital forensics and computer security. He is also co-founder of Digital Forensics Solutions, a private digital forensics corporation, and a technical advisor to the Gulf Coast Computer Forensics Laboratory (GCCFL).

VASSIL ROUSSEV (Vassil@roussev.net) is an assistant professor in the Department of Computer Science at the University of New Orleans.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.