World Scientific and Engineering Academy and Society (WSEAS) Transactions on Computers, Issue 5, Volume 3,
ISSN 1109-2750, pp. 1295-1298, November 2004.
Proceedings of the 4th WSEAS International Conference on Information Science and Applications,
ISBN: 960-8052-97-1, April 21-23, 2004.

A Virtual Test Bed for Distributed Processing of Archives

BINH NGUYEN
U.S. Army Research Laboratory
Adelphi, MD 20783
USA
bnguyen@arl.army.mil

Abstract: - Performing empirical research requiring numerous interactions with web servers and transfers of very large archival data files without affecting operational information system infrastructure is highly desirable by separating unsteady test-bed environments from steady administrative networks. The key theme of this paper is the building of a virtual network of heterogeneous computing systems using virtual machine technologies. The virtual network exists to provide a low-cost and convenient environment for conducting applied research in support of operational requirements for the secure transfer and storage of distributed archival electronic records over a public network, such as the Internet. First, the paper describes a virtual network environment in which security technologies and methods are being developed, tested, and evaluated. Second, it discusses the concepts and methods for building such a virtual system. The paper then concludes with positive preliminary results and suggestions for processing electronic records archives using virtual machine technologies.

Key-Words: - Virtual machines, simulated computer networks, virtual test bed

1 Introduction

Archival data files used in this paper refer to computer-generated ordered collections of related files that need to be authentically preserved. For a computer user, an archival data file is often merely a compressed or uncompressed archive that consists of one or more files. An archive is often in a form of a *tar* file, a *zip* file, a *jar* file, or a self-extracting *zip* file, depending on the employed archival tool in the process. Examples of data files include databases, spreadsheets, email, architectural and engineering drawings, artifacts, maps and charts, moving pictures, photographs, sound recordings, or textual records [1].

Distributed processing of archival data files with limited budgets requires not only an affordable computing environment, but also a secure access to the source of the data files and a convenient means to exchange information among geographically dispersed collaborative researchers. An approach to connect remote processing sites is to use web technologies and the Internet because they are a relatively lower-cost alternative than other kinds of private networks, including leased or privately own telephone lines. Popular web technologies include web servers and browsers that use standard cryptographic algorithms and protocols capable of providing the stakeholders of archival data files with reasonable information assurance services, including authentication, confidentiality, integrity, and non-repudiation.

Building a prototypal web-portal for distributed archival data processing over the Internet is a workable solution for gauging the performance of the potentially employable functional and security technologies in an operational environment in the future. However, doing so would require substantial commitment of financial resources and the involvement of geographically distant personnel. In addition, frequent testing and evaluation of the portal require interactions with its web servers and transfers of very large archival data files can potentially create severe network contention and congestion on local administrative networks, internet service providers, and other interconnected network computing systems.

An alternative solution is to build a functionally equivalent prototype using software emulation of a physical computing and networking environment using virtual machine technologies. This solution requires as few as one personal computer operated by one person. All computational processes and communications traffic occur within a physical computer, and thus does not affect any other computing networks. Software emulation tools that enable the building of virtual machines include, but are not limited to, *VMware*[1] and *Virtual PC*[2].

[1] http://www.vmware.com
[2] http://www.microsoft.com/windowsxp/virtualpc

The next section of the paper delves into the details of the virtual machine technologies and provides some references to other work that also use virtual machines to build virtual network and computing infrastructure for research and for educational purposes. Section 3 explains the employed method for building the virtual network test bed. Section 4 presents some preliminary results and discusses their implication. Section 5 describes future empirical research efforts, summarizes the benefits of using virtual machines, and concludes the paper with some recommendations for processing archival data files in a virtual environment.

2 Virtual Machine

Virtual machine is a software application that emulates a real physical machine. It is an application in a sense that it is controlled by the host operating system that manages all the computing resources in a real machine. The virtual machine itself is not useful until it runs a guest operating system and its applications. Current virtual machine technologies enable the building of not only virtual machines, but also virtual networks to which virtual machines are connected. Thus, a portable notebook computer can host a complete virtual network of virtual machines running various appropriate operating systems.

Each virtual machine has the same system requirements as a real physical machine because each virtual machine is a realistic emulation of an actual machine. For example, if a minimally configured machine needs 128 MB of memory and 4 GB of hard drive to function properly, then a computer that hosts N virtual machines running simultaneously would need an additional N times the resources needed by each virtual machine. Moreover, running multiple virtual machines requires a faster processor because virtual machine technology is a software emulation. In general, the faster the processor speed and the more memory the host computer has, the better the performance the virtual machine delivers.

Virtual machine technologies are not new, but they were not popular until these days when the cost-performance ratio of computing technologies is substantially reduced. According to a research by McEwan [2], the International Business Machine (IBM) Corporation introduced the first virtual machine about 40 years ago to run on its mainframe computers, and now interest in virtual machine is growing among commercial vendors, open-source community, and academic research institutions as the power of personal computers continues to grow.

A sample of applications of virtual machine technology includes the simulation of old machines [3], grid computing [4], and research and educational purposes [5-7]. Following paragraphs summarize the last application category. The Christchurch Polytechnic Institute of Technology (CPIT) uses virtual machine technology to build network and computing facilities for education in areas of information and communication technology [2].

The United States Military Academy (USMA) and the George Washington University (GWU) use virtual machine technologies to build hands-on laboratories for educational and research purposes. The USMA laboratory is for information warfare analysis and research [5,6] and the GWU laboratory is for computer security and information assurance educational purposes [7]. The USMA and the GWU provide their students with realistic network environments in which the students could launch simulated cyber attacks and learn how to defend their networked automated information systems.

3 Method

A mix of commercially available and open system software systems and a notebook computer running The *Microsoft Windows XP Pro* operating system facilitate the building of the virtual test bed. The notebook computer is a *Dell Inspiron 8500* equipped with an *Intel Pentium 4* microprocessor, 2 GB of memory, and a 40 GB hard drive. The virtual machine technology is the *VMware Workstation 4.0* that enables the construction of several virtual machines running various operating systems and their connection to a virtual network.

Disparate operating systems running on the virtual machines include *Red Hat Linux*, *Microsoft Windows XP Pro,* and *Microsoft Windows 2000* operating systems. Each virtual machine is configured with a minimum of 128 MB of main memory and a 4 GB hard drive. The size of the main memory of a virtual machine is reconfigurable, but the size of its virtual hard drive is not. The heterogeneity of the installed operating systems reflects realistic distributed processing environments in which dissimilar computing systems used by various researchers located at different institutions having disparate computing systems.

Initial network topology is a virtual local network configured as a host-only networking environment, a special network option of the *VMware Workstation* software. This setup creates a private network to which all the virtual machines and the host machine connect by means of virtual network adapters and a virtual switch. Each adapter has a manually assigned unique, permanent private internet protocol address [8]. This arrangement provides an isolated virtual network and enables the virtual machines to have a direct connection with the host operating system running in the real machine.

Preliminary steps performed to accomplish two objectives: (1) to substantiate the connectivity among the virtual and the real machine included the installation, and (2) to experiment with network services programs such as file transfer and web services. Tested file transfer services consists of mainly unsecured file transfer protocol (*ftp*) and secured copy (*scp*) commands available in the Linux operating system. Tested web services employed unsecured and secured hypertext transfer protocols (*http* and *https*), respectively. Nearly all tested communication scenarios involved a server and several clients. The server was always a virtual machine, and the clients were other virtual machines and the real machine.

4 Results and Discussions

Although this is the first time that a virtual network of virtual machines was set up on a notebook computer, the installation process went relatively smoothly. Prerequisites for the set up required sufficient procedural and operational knowledge of the virtual machine, various system requirements and idiosyncrasies of each guest operating system, and the intricacies of internetworking technologies.

Experimentation with network services running in virtual machines required configuration and testing just like doing so on a real machine. This is obvious in a sense that the network applications were running on a real operating system, not a virtual or a simulated one.

Performance deteriorated when more than four virtual machines were running concurrently, but it was acceptably reasonable. A way to improve the performance was to add additional physical memory to the real machine and then increased the amount of main memory available to each machine.

The virtual test bed uses the *VMware* software because of three reasons. First, the software supports several dissimilar operating systems; among them are the two popular operating systems, the Microsoft Windows and the Linux operating systems. Second, it is relatively mature and stable product. Third, its vendor provides technical support for the product.

The reason for using notebook instead of a desktop to run the virtual test bed was mainly for its portability. Portability facilitates the demonstration of concept and a prototype of the system to potential clients and supporters at various geographically different locations.

The set up of the virtual test bed could have been easier and faster with the installation of a single, homogeneous operating system. However, doing so would be unrealistic because the anticipated distributed environment for processing electronic records archives is a network-computing environment in which diverse computing systems are interconnected.

The virtual test bed is also potentially a suitable environment for processing two special types of electronic records archives: (1) archives that can function only in an older operating system, and (2) archives that cannot be trusted to run in a physical machine. An example of the former type is a database management system and its associated data files taken from a computer that ran the Microsoft Disk Operating System (DOS) in the 1980's. A virtual machine can accommodate an old operating system such DOS, which in turn provides an appropriate environment for legacy applications. An example of the latter is an executable archive or an archive that contains executable files whose functionality is completely unknown to a new user. A misbehaving executable file can cause damages only to the virtual machine. Replacing or rebuilding a virtual machine is easier and less costly than a physical machine.

4 Summary and Future Work

This paper has described how modern virtual machine technologies can enable the building of a virtual prototype of a virtual network environment. The reason this test bed exists is to conduct empirical experimentations of defensive information assurance technologies capable of securing and protecting sensitive electronic records archives that will be stored, processed, and transmitted in an unsecured public network. The ability of virtual machines to realistically rendering a physical machine enables the building of a heterogeneous

computing environment with much lower costs and convenience.

The isolated virtual network enables the realization of three important benefits: (1) the complete control of all types of network traffic generated in the test bed, (2) the facilitation of system performance evaluation and measurement of security overhead, and (3) the complete independence from organizational administrative networks and their administrators.

The next step in this effort will be the generation of simulated electronic records archives and the building of a prototypal web-portal capable of operating in unsecured and secured mode. Other efforts will also involve the evaluation of some defensive security products that are deemed highly appropriate and effective for the protection of the computing systems and networks that store, transmit, and process sensitive electronic records archives.

Additional effort for experimentation with network routing protocols and performance evaluation will include the establishment of a virtual interconnected network of several sub-networks. Each sub-network is a separate domain having different functionality.

5 Conclusion

Although the virtual test bed was built to perform empirical experimentation of security technologies, it can be easily amenable to host actual processing of sensitive electronic archives. The isolation of the test bed from the administrative network provided an ideal environment in which untrusted executable files can be executed for the inspection and evaluation of its behavior. The virtual test bed also offers a practical networked computing environment for educational and empirical research activities.

Acknowledgements

The author sincerely appreciates the support for this work from the U.S. National Archives and Records Administration and the constructive comments of anonymous reviewers.

Disclaimer

The findings in this paper are not to be construed as an official Department of the Army position unless so designated by other authorized documents. Citation of manufacturer's trade names does not constitute an official endorsement or approval of the use thereof.

References:

[1] National Archives and Records Administration, AAD [Access to Archival Databases] Terminology, URL: http://www.archives.gov/aad/terminology.html (accessed 20 Nov 03)

[2] W. McEwan, Technology and Their Application In The Delivery of ICT, *Proceedings of the 15th Annual Conference of the National Advisory Committee on Computing Qualifications*, 2nd - 5th July 2002 Hamilton, New Zealand. URL: http://site.tekotago.ac.nz/staticdata/papers02/papers/mcewan55.pdf (accessed 21 Nov 2003)

[3] B. Supnik, The Computer History Simulation Project (Referenced in [2]), URL: http://simh.trailing-edge.com/ (accessed 10 Dec 03)

[4] R. Figueiredo, P. Dinda, J. Fortes, A Case For Grid Computing on Virtual Machines, *Proceedings of the 23rd Internatinal Conference on Distributed Computing Systems (ICDCS),* May 2003. URL: citeseer.nj.nec.com/figueiredo03case.html (accessed 19 Nov 2003)

[5] D. Ragsdale, S. Lathrop, R. Dodge, Enhancing Information Warfare Education Through the Use of Virtual and Isolated Networks, *Journal of Information Warfare*, Vol. 2, Issue 3, pp. 47-59. http://www.itoc.usma.edu/Documents/JIW_rags.pdf (accessed 20 Nov 2003)

[6] R. Dodge (MAJ), D. Ragsdale (LTC), and D. Welch (COL), State-of-the-Art Information Warfare (IW) Training, *IA*newsletter, Volume 6 Number 2, Summer 2003, Information Assurance Technology Analysis Center, Falls Church, VA 22042, USA, pp. 18-19,28-29.

[7] L. Hoffman, T. Rosenberg, S. Willmore, The Portable Educational Network (PEN), Computer Science Department, The George Washington University, Washington, DC 20052, URL: http://www.cs.seas.gwu.edu/seccert/pen.doc (accessed 20 Nov 2003)

[8] Y. Rehhter, B. Moskowitz, D. Karrenberg, G. J. de Groot, and E. Lear, Address Allocation for Private Internets (RFC 1918), February 1996, URL: http://www.ietf.org/rfc/rfc1918.txt (accessed 11 Dec 03)