Decentralized Virtualization in Systems Administration Education

Bill Stackpole
Department of NSSA
Rochester Institute of Technology
Rochester, NY 14623
Bill.Stackpole@rit.edu

Jason Koppe
Department of NSSA
Rochester Institute of Technology
Rochester, NY 14623
jrk9055@rit.edu

Tom Haskell
Department of NSSA
Rochester Institute of Technology
Rochester, NY 14623
tbh1394@rit.edu

Laura Guay
Department of NSSA
Rochester Institute of Technology
Rochester, NY 14623
lrg3267@rit.edu

Yin Pan
Department of NSSA
Rochester Institute of Technology
Rochester, NY 14623
Yin.Pan@rit.edu

ABSTRACT
The Networking and Systems Administration Laboratories are vital components of the applied academic experience of the students of the Networking, Security and Systems Administration Department. Students are currently able to implement multiple heterogeneous computing infrastructures using Symantec’s Ghost Solution Suite. This paper presents and explores virtual lab environments that support alternative preparation solutions that will increase laboratory utilization, student productivity, and provide students with exposure to virtualization technology.

Categories and Subject Descriptors
K.3.2 [Computer and Information Science Education]: Computer science education, information systems education

General Terms

Keywords
Labs, virtual machines, VMware, virtualization, linked clones, system administration, infrastructure, storage, scalable, performance, lab management

1. INTRODUCTION
Since the inception of system administration education, universities have attempted to build laboratories where students could apply system administration concepts in a practical and cost-effective manner. The problems with proposed and implemented laboratory architectures for system administration education include time-consuming management, low mobility, and inefficient resource utilization. Various studies have included performance and cost analyses regarding the use of virtualization technologies in combination with storage systems to deliver manageable virtual infrastructures to students in system administration courses. Most of these studies involved lab environments that are smaller than the RIT system administration lab; further, these studies do not support the heterogeneous architectures that RIT requires. The lack of qualitative analysis and quantitative support regarding the use of decentralized virtualization platforms to facilitate scalable heterogeneous educational system administration laboratories means that educators have not yet realized or implemented a viable solution to the problems.

The Networking and Systems Administration Laboratories are vital components of the applied academic experience of the students in the Networking, Security, and Systems Administration Department. Currently, students are required to implement between one and ten systems using Linux, UNIX and Microsoft Windows clients and servers within multiple networks. The lab is home to 80 student workstations, each with Intel 3.4 GHz processors, 1GB of RAM, and two network adapters (NIC). During their laboratory period students use Symantec Ghost to save and restore custom operating system implementations; a one-gigabit NIC is dedicated for this type of student use. The other NIC is connected to a physical hub so that students can interrogate network packets.

The laboratory for system administration is nearing one hundred percent utilization. In addition, preparation and teardown can consume up to half of the laboratory period. Thus, the imaging infrastructure is becoming increasingly cumbersome.

The primary goals of this study include identifying problems contributing to lab inefficiency and surveying technologies to address the problems. After demonstrations, feasibility tests, undergraduate research, related infrastructure management, and pilot tests [5], this paper presents analysis of infrastructures that utilize VMware’s virtualization platform to enable dynamic, low-cost virtual environments that students can control in traditional
physical laboratory space. Students who spend less time preparing and saving their infrastructures spend more time learning.

2. RELATED WORK
Educators and researchers [1, 2, 3, and 4] described difficulties creating a computing environment which students are required to control for system administration education.

Begnum, Koymans, Krap, and Sechrest (2004) described the use of User-Mode Linux (UML) as a virtualization platform and My Linux Network (MLN) as a virtualization administration tool at university and industry environments. Begnum et al. concluded that User-Mode Linux enabled students more efficiently learn system administration concepts. The authors stated that their approach to enabling students to apply system administration concepts need only function “as specified in the RFC’s” – therefore, they were not required to offer specific operating systems or applications, just something that “worked correctly.” While their UML/MLN architecture enables system administration education in their institution, RIT requires implementation of heterogeneous architectures including non-Linux operating systems.

Stockman, Nyland, and Weed (2005) presented findings regarding a centralized delivery of virtual machines to a small lab environment. The authors concluded that mobility could be achieved with a central storage for student systems; however, Stockman [6] stated that the centralized delivery was never implemented because the storage system did not scale as expected. The use of Microsoft Virtual PC as a virtualization platform limits types of supported virtual operating systems to only those from Microsoft; RIT requires implementation of heterogeneous infrastructures that include non-Microsoft operating systems.

Stockman et al. did not describe usage of Virtual PC’s differencing disks. The study offered by Vollrath and Jenkins (2004) explored the implications and cost of using Virtual PC to alleviate space and sign-off problems. The use of virtual machine differences (analogous to VMware’s linked clones) is a novel approach to lower storage resource cost. However, Virtual PC [7] is not the right solution for RIT due to the lack of support for non-Microsoft operating systems.

In 2003, Lei and Rawles surveyed performance and cost of three virtualization technologies that would enable a more practical lab environment. Their research included both quantitative methods involving performance benchmarking with different storage and virtualization technologies and qualitative methods regarding cost analysis. The results concluded that any virtualization technology implemented because the storage system did not scale as expected. The use of Microsoft Virtual PC as a virtualization platform limits types of supported virtual operating systems to only those from Microsoft; RIT requires implementation of heterogeneous infrastructures that include non-Microsoft operating systems.

In 2003, Lei and Rawles surveyed performance and cost of three virtualization technologies that would enable a more practical lab environment. Their research included both quantitative methods involving performance benchmarking with different storage and virtualization technologies and qualitative methods regarding cost analysis. The results concluded that any virtualization technology coupled with a Microsoft Windows host operating system and a networked storage system provided the most cost-effective and highest performing environment to enable applied system and networking administration learning. This study is unique in that it quantifiably measured performance of the system at multiple points: host resource utilization, virtual OS installation time, and network utilization. Future work should revisit the quantitative analysis methods of Lei and Rawles to realize performance and cost transformations since their study.

3. BACKGROUND & REQUIREMENTS
Traditional lecture courses offered from the department of Networking, Security, and Systems Administration (NSSA) are complemented with hands-on lab exercises. These exercises require students to apply concepts learned during the lecture to a computing environment that they design, configure, and manage.

The labs were created to provide an environment suitable to conduct lectures, labs and open hours. Open hours refers to a period of time when there is no scheduled lab or lecture and students are able to work freely. Further, the labs were designed with the intent that, prior to the start of the lab session, the lab environment should be prepared to save the student from this task. The designers understood that experimentation takes time and they wanted to allow students to save their progress so they could return to complete their exercises later. Coupled with Active Directory (AD) authentication and granular access control, the designers enabled the lab to support student system backup and restoration. Over the years, strategies to boot to a Ghost-capable platform included floppy disks, compact disks, and most recently a diskless pre-boot execution environment (PXE). The Ghost-capable platforms include DOS relatives like MS-DOS, PCDOS, FreeDOS, etc. These technologies make possible pre-lab imaging and student managed imaging.

A considerable amount of work and money has gone into creating the existing laboratory infrastructure. Cost effective solutions should embrace and extend existing infrastructure. Therefore, utilizing existing infrastructure is a requirement.

4. PROBLEMS
There are four major problems to be addressed: (1) imaging time (2) manageability of imaging infrastructure (3) inefficient lab space utilization and (4) computing resource utilization. These problems share some of the same causes.

4.1 Imaging Time
The amount of time it takes to save or restore all the machines in the laboratory has become unacceptable.

Simultaneously restoring all eighty machines in the system administration lab was initially a five to ten minute process. However, with the average size of an operating system disk image growing, the restoration time has increased accordingly. It now takes up to twenty minutes to re-image a machine. When the lab is being prepared in advance of student arrival, this increase in imaging time is not a problem. However, during the course of the academic quarter it becomes necessary for students to store and retrieve their customized system images after their arrival at the laboratory. These two processes (store and retrieve) require twenty minutes each. Thus, the productive time available for each lab period is reduced from one hundred and ten minutes (one hour and fifty minutes) to only seventy minutes.

When forty students re-image all eighty machines, they are typically transferring between one (1) and ten (10) base operating systems in order to complete their lab exercises. These base operating system images are similar for all twenty benches in the lab – with differences being limited to small variations in configuration information.
Imagine two students working on the same lab assignment. The assignment requires the students to maintain an environment consisting of a Windows XP system and a Windows 2003 system. The students must both copy the same data for Windows XP across the network to two separate machines as well as copy the same data for Windows 2003 across the network to two other machines. A similar scenario exists when the students need to stop and save their progress. While the students may have slightly different implementations of the assigned environment, much of the underlying data being copied to the file servers will be the same. This process repeats itself throughout the each academic session.

Given the limited changes made by lab users and the similarities in base operating system images, there is redundant network traffic. When the labs are being prepared (either in advance or during the lab period), duplicate instances of complete operating systems are being copied across the network. The overhead caused by repeatedly copying the same underlying operating system across the network can be considered unnecessary. It adds to the preparation time and consumes bandwidth that could be put to better use.

4.2 Manageability
The periodic refresh of computing and infrastructure hardware requires the operating system images be refreshed as well. Refreshing these images provides hardware and OS compatibility but is a time-consuming process.

The labs in the infrastructure are each categorized in terms of the maximum age and minimum performance of the hardware needed to support them. Labs with the highest performance requirements receive new hardware each refresh period while each of the labs ranked lower receive “hand-me-downs” from the labs above them. This aptly named “cascade process” has proven effective at reducing capital expenditures required to keep the labs up to date. With respect to the management of the lab, the cascade process brings a unique set of problems.

The labs participating in the cascade do not share identical infrastructure and hardware. The variety of hardware platforms require different sets of images to be created to support them. Each of the five labs in the cascade requires a complete rebuild of each of the base operating system images to coincide with the hardware changes in their physical infrastructure. For example, there are five different copies of the same Windows XP image – one for each physical lab. New base images are created every time a cascade takes place.

The cascade process builds into the lab infrastructure an inconsistency of physical hardware platform when moving from lab to lab. While every effort is made to ensure that future needs are considered, the cascade process does not guarantee that there will be enough of a given asset to supply the labs. These issues could be due to equipment failure, inconsistency of hardware with current or older operating systems, etc. Efficient lab management should work to provide consistent hardware architectures shared across all the labs.

4.3 Lab Space Utilization
Unique hardware configuration in the labs reduces lab space utilization and efficiency.

Since each of the five labs was tailored to support a different facet of computing, the current lab spaces are sometimes seen as inflexible. This can result in physical lab space being used only for lab exercises specific to the course for which the lab was designed.

The labs originally incorporated variation in their hardware configurations to ensure they effectively supported specific courses. For instance, exercises in system administration often require the use of multiple subnets. To that end, the System Administration lab (SysLab) was built with 80 different useable subnets as a feature to allow any bench to be served by up to four different subnets simultaneously. Since many classes require students to configure systems on multiple subnets and the SysLab is highly flexible in this regard, it tends to be more highly utilized than some of the other labs. This utilization nears one hundred percent at the end of the quarter, while other labs, which do not share this feature of their infrastructure, are often under-utilized by comparison. The goal of providing consistency in lab capability while maintaining flexibility in the infrastructure should result in greater scheduling flexibility, more consistent utilization, and greater operational efficiency.

4.4 Computing Resource Utilization
The current “one machine per operating system” model results in high-performance machines being under-utilized and relegated to perform low-performance tasks.

Given the tasks required to demonstrate capability in the Network- and Systems-Administration classes, the cascade process provides each of the labs with higher-performance platforms than are absolutely required to support these endeavors. As a result, the machines in the lab are capable of supporting more load than they are currently required to support. A reasonable expectation would be to exercise the hardware to a greater degree, leveraging the performance available on the platform. This could help relieve the bottlenecks experienced at the end of the quarter by reducing the number of machines required per student and thus increasing the student density in the labs during open lab hours.

For example, a student in the Network Services course must run DNS server on one physical computer and perform DNS queries from another. Consolidating these roles so that they all function from the same physical machine more effectively utilizes the physical hardware.

Another issue being faced is a perceived difficulty in obtaining Ethernet hubs. Due to the continued focus on security in the computing arena, it appears that the hardware manufacturers are building more unmanaged switches and less “simple” repeaters and hubs. Newer devices are less apt to broadcast all traffic over all connected ports. This prevents students from seeing the packet traffic and limits their ability to interpret the traffic on the network. While simple repeaters are more difficult to locate and purchase, certain networking hardware can perform port mirroring, which can assist in packet capturing; however, the solutions outlined in the following section lessen the overall dependency on physical networking devices without increasing monetary expenditures.
5. SOLUTIONS
Virtualization is the solution for the previously addressed problems. Virtualization eliminates student issues with hardware specific images using the hardware abstraction layer; this means that virtualization supports mobility, more efficient use of lab space and lower management costs. Further, virtualization enables efficient utilization of computing resources by executing multiple virtual operating systems on each physical host. The requirement for students to save and restore work is achieved through the centralized storage of linked clones.

While virtualization is the proposed solution, technology and management variables are addressed in this section.

5.1 Virtualization
A variety of virtualization platforms were considered including User-Mode Linux (UML), Microsoft Virtual PC, VMware Workstation, and Xen. Virtualization eliminates student hardware-dependent images, addressing the problem discussed in section 4.2. Decentralized virtualization most effectively utilizes the computing power present in the existing infrastructure. While UML, Microsoft Virtual PC, and Xen might have offered features similar to VMware’s linked clones, VMware offered the most guest operating system support, the highest cross-platform compatibility and required no software or hardware purchases (see section six and [8]).

The use of VMware’s linked clones is critical to the efficient use of network and storage resources. VMware defines a linked clone as “a copy of a virtual machine that shares virtual disks with the parent virtual machine in an ongoing manner […] while […] changes to the disk of the linked clone do not affect the parent.” While the size of a modified operating system image is the sum of the size of the operating system and the modifications, the size of a linked clone is merely the size of the modifications. When saving modifications, linked clones consume less storage space; therefore, linked clones more efficiently use disk space than full system images.

5.2 Storage
The parent virtual machines should be stored on every workstation to resolve the problem of transferring redundant operating system images during lab sessions (see section 5.5 for workstation management).

The linked clone should reside within the shared storage system to support mobility. Shared storage solutions that have been considered include Gmail, personal storage devices, networked file servers, and storage area networks. A single, centralized storage area network is easier to manage than a distributed model.

5.3 Execution
The linked clones can be executed either locally or remotely. Local execution requires copying the linked clone to the client workstation and then executing the virtual machine. In a remote execution, there is no need to transfer the linked clone; it remains within the shared storage.

Remote execution supports the ultimate goal of instantaneous lab preparation. The remote execution of linked clones may have a performance cost to both the storage and the network that is believed to be insignificant and will be addressed in future work.

5.4 Network
The network must support large data transfers between the shared storage and each client. As mentioned previously, the current systems have gigabit network connections that provide this capability.

To achieve lab-to-lab mobility, alternate network architectures should be considered, specifically ones that support consistent access to the SAN regardless of client location.

5.5 Management
To maintain consistency and leverage the strengths of the existing AD infrastructure, the client workstations in the lab will be joined to the domain. Microsoft Windows leverages AD integration, while other operating platforms do not.

The host operating systems must still interact with physical hardware. To that end, deployment of host machine images must address hardware dependencies. Microsoft’s Windows Deployment Services (WDS) enables multicast installation of customized Windows operating systems to a variety of physical hardware configurations. Further, WDS can automatically rename and join those installations to an AD domain. Linux does not appear to offer system deployment solutions as robust and tightly integrated as WDS.

Managing parent virtual machines at the workstation level is required to eliminate the problem of transferring duplicate data. In order to provide mobility and a consistent user experience, standardization of parent virtual machines is required. Initially, parent virtual machines deployed within the file system of the host machine image are standardized. However, assignment and management adjustments periodically require modifications or additions to the initial set of standard parent virtual machines. Three (3) options offer differential updates for parent virtual machine management while a fourth, utilizing WDS, provides a complete re-deployment of the host machine image and its contents. The three (3) differential update options include:

1. Pushing updates for parent virtual machines from shared storage (e.g. using robocopy [11])
2. Pulling updates for parent virtual machines from each client workstation
3. VMware ACE [10]

The most appropriate update option has yet to be determined.

6. COST ANALYSIS
For the implementation described, there is zero additional monetary cost because the discussed infrastructure, hardware, and software have already been acquired. The cost to obtain Microsoft Windows for servers is the highest. Similar to MSDNAA [12], the VMware Academic Program [8] provides free use of appropriate VMware products in academic instruction or labs.
7. CONCLUSION / FUTURE WORK
Considering the problems system administration educators are faced with and the solutions discussed in this paper, distributed storage of parent virtual machines with centralized storage of linked clones provides a manageable solution that is cost-effective, efficient, and scalable. Decentralized virtualization most effectively utilizes the computing power present in the existing infrastructure. VMware’s Workstation product supports a variety of virtual operating systems. The Microsoft products provide authentication, access control, deployment and feature integration at the workstation. The inclusion of a storage area network provides a highly available location for students to store their progress. Finally, the use of VMware’s linked clones in combination with these evaluated technologies decreases the requirement for student-specific data transfer and archival; thus making this environment scalable.

Future work will implement the systems described as well as provide quantitative analysis of proposed systems, investigate increasing the mobility of students from lab-to-lab, and consider providing synchronization of student work with a remotely accessible environment.

8. ACKNOWLEDGMENTS
We extend gratitude to all of the faculty and staff at RIT as well as students Ron Valente and Ben Allen for all of their help and support with these solutions.

9. REFERENCES


