Virtualization, Application Streaming and Private Cloud Computing in a Training Laboratory

Jiabin Deng*
Computer Engineering Department, Zhongshan Polytechnic, Zhongshan, China
Email: hugodunne@yahoo.com.cn

Juanli Hu and Anthony Chak Ming Liu
Computer Engineering Department, Zhongshan Polytechnic, Zhongshan, China
Email: hjlfoxes@163.com and acmliu@hotmail.com

Abstract — In a typical I.T. training laboratory software applications and operating system are both installed onto each workstation. Its configuration management (server and workstation provisioning, de-commissioning and change control) often consumes a vast amount of resources such as time and human resources. Also, in a poorly managed lab environment, configuration management and its functional efficiency are usually the least to be looked at or measured upon. Systems virtualization and applications streaming thus provide a way to streamline the process of configuration management. Virtualization is designed to achieve maximum utilization of system resources such as network bandwidth, storage, CPU and RAM. In a private cloud, the operating system and applications are stored centrally and then streamed to the workstations, on-demand and in real time, but all processing is done behind the scene within the cloud infrastructure. The objective of this paper is to illustrate the key steps our project team will undertake to build a Private Cloud in order to improve upon the efficiency of configuration management within our I.T. training laboratories.


I. INTRODUCTION

Technicians who are working in a busy training laboratory often find them pressurized, battling against time, unrealistic deadlines and numerous change control as well as configuration management to the I.T. systems. With the maturity of virtualization [2] technology and affordability they become a reality for any academic institution to build its own data centre. Streaming a Windows™ software application into an isolated environment effectively accelerates I.T. deliveries by reducing regression testing [6] and simplifies management with streamlined maintenance, upgrades and de-provisioning. It has been suggested that using streaming for de-provisioning an application is probably the most efficient method of removing all traces of an application. Virtualization model and its layers [2] help to improve work efficiency by leveraging huge computing ability and storage resource in a private cloud [1] [4].

Virtualization and Private Cloud Computing are both relatively new in China. Our project team undertakes the challenge to build a Cloud infrastructure which is capable of accommodating more than 2500 desktops spread between 40 training laboratories, excluding 1000 PCs used by staffs and lecturers. There are a number of technical objectives our team intends to achieve. They are:

1) on-demand application streaming.
2) dynamic server provisioning.
3) centralized access to the presentation layer.

II. FRAMEWORK

A. Virtualization Model

The virtualization model, in Fig. 1, illustrates five key layers which will be built in our project. Each layer forms its own virtualized resource pool.

The first layer is Network Virtualization. It is hardware and software technology which presents a view of the network that differs from the physical view. For one instance, a PC may be allowed to only connect to systems it is permitted to access, where another instance is making multiple network links appear to be from a single link.

The second layer is Storage Virtualization. Again, it is hardware as well as software technology which conceals where storage systems are and what type of device is actually storing applications and data. New storage technology makes it possible for many systems to share the same storage devices without knowing that others are

*Corresponding author.
also accessing them. In addition, the same technology also makes it feasible to take a snapshot of a live system so that it can be backed up without intruding online or transactional applications.

Third, the layer is Server Virtualization. It is about hardware and software technology that conceals physical hardware configuration from operating systems, system services or applications. The technology makes one system appear to be many, or many systems appear to be originated from a single computing resource. There are a range of objectives which needs to be achieved and they are: raw performance; high levels of scalability; reliability; availability; agility or consolidation of multiple environments onto a single system.

Fourth, Application Virtualization is about software technology which lets applications to run on many different operating systems and hardware platforms. Some of the more advanced forms of this technology offer the ability to restart an application in case of a failure; start another instance of an application if the application is not meeting service level objectives; or provide load balancing among multiple instances of an application in order to archive high levels of scalability.

Fifth, the Presentation Virtualization is about hardware and software technology which lets almost any device to access any application without having to know too much about the other. The application “sees” a device it is used to working with. When the device “sees” an application it knows how to display. In some occasions, special purpose hardware is used on each side of the network connection in order to increase performance; allow many users to share a single client system or allow a single individual to see multiple displays.

As with any system, access, security and system management must also be considered. In a virtualized environment it is about software technology that controls access to systems, users and file level security arrangements, and management which makes it possible for multiple systems to be provisioned and managed as if they were a single computing resource.

Upon successful implementation of all five layers a Cloud infrastructure is formed and it is ready for the implementation of application streaming.

B. Virtualized Infrastructure – A Walkthrough

Building a virtualized infrastructure as demonstrated by VMWare [8] in Fig. 2, is about separating the software environment from its underlying hardware infrastructure so multiple servers, storage and networks are placed into a shared pool of resources. These resources are then dynamically delivered securely and reliably to applications where and when needed. This approach allows technicians use blocks of industry-standard servers to build a self-optimizing Data Centre and deliver high levels of flexibility, scalability, utilization, availability and automation.

A virtualized network is achieved when these conditions are met:
1) multiple network links appear to be a single link. It is about the combined network utilization of all virtual machines (VM) from the same physical host server.
2) Using a Gigabit network and equipments to increase network throughput.

Virtualized storage can be achieved when many systems sharing the same storage devices without knowing that others are also accessing them. Virtualized storage also offers the possibility of taking a snapshot of a live system so that it can be backed up without intruding online or transactional applications. To meet these objectives in the commercial world, Storage Attached Network, or SANs, is widely adapted.
The physical host server, waiting to be virtualized, is just like any industry standard server. Most server configurations share some of these common characteristics: multi-core processors, Gigabit network card, a large quantity of RAM (e.g. 128 Gb or above), ultra fast hard drives and etc. Before the virtualization process begins, a piece of software called the hypervisor which needs to be installed onto the host server in addition to its operating system.

A hypervisor, also known as a Virtual Machine Monitor (VMM), is a software program which allows multiple operating systems to share a single hardware host. Each piece of operating system appears to share the host's processor, memory, and other resources. In addition, one of the functions of hypervisor is actually controlling the host processor and resources; allocating what is needed to each operating system in sequence and making sure that the guest operating systems cannot disrupt each other.

VMM divides the hardware host’s multi-core processors into a number of Virtual Machines (VM). The hypervisor configures and scales each VM with a number of Virtual CPU (vCPU); allocate memory to each VM; and support I/O intensive applications such as SAP, SQL Databases, Microsoft Exchange and others.

VMware claimed that VMM to be one of the best ways leveraging the capacity of the latest multi-core processors found on many server hardware hosts. Each VM has its own Operating System and can run applications independently from within its own virtual environment.

In a Data Centre environment it is a common industry practice that servers are placed into groups called cluster nodes. Inside a cluster, policy to cater for the host server failover is configured. If and when a server fails another “parked” server inside the same cluster will automatically take over. This process is entirely automatic by setting up relevant policy in the management console.

III. PROPOSED SOLUTIONS

A. Application Streaming

In a conventional x86 computing model the applications are tightly coupled to physical servers. The model is too static and fragmented to efficiently support today’s complex and dynamic applications.

Application virtualization dynamically delivers applications by creating a pre-configured application image that can be provisioned on-demand. With the help of the sequencing and packaging software, standard and pre-configured application stacks can be deployed via the VMM Console.

The process involved, as demonstrated by Microsoft™ [9], is called Sequencing and Packaging [3]. Sequencing is the executing or running the application in a way that all the streaming blocks are optimized before delivering to the client.

Packaging is the creation a set of image and configurations files that will later be installed automatically on a client. Once an application is sequenced and packaged, the files are usually stored on a streaming server. Sequencing involves 3 phases: Installation, Configuration and Execution. The installation phase involves the installation of applications. The configuration phase involves processing as if it were a local application. The execution phase performs some of the same functions that users will be likely to perform.

Fig. 3 illustrated the application sequencing and packaging workflow. The two output files are: a header file and a stream file. The stream file contains one or more sequenced applications that the sequencer has packaged into streaming blocks plus the delivery information.
Application streaming is an on-demand distribution for non-web-based software. The basic concept is that only particular parts of a computer program need to be available at any given time for the end user to perform a particular function. Hence, a program need not be fully installed on a client computer, but part of it can be delivered over a low bandwidth network as and when they are required.

The streaming service enables applications to be delivered to client devices and run in protected and isolated virtual environments. Packaged applications are centrally managed in a streaming cluster, and are then streamed to the client device. These packaged applications then become an on-demand service that is always available and up-to-date. Application caching technology makes applications available even when users are not connected to the network.

There are two effective ways of streaming namely Push & Pull. Pulling happens when a request was originated from the end user; the client is pulling the application from the streaming server. Pushing occurs when the application might be delivered from the streaming server to the client in background.

B. Network Load Balance & Clustering

Network Load Balancing (NLB) is implemented in a special driver installed on each host server in a cluster. The cluster presents a single IP address to clients. When client requests are received, they go to all host servers in the same cluster, and an algorithm implemented in the driver maps each request to a particular host. The other hosts in the cluster drop the request. Load partitioning can be set, via the management console, to distribute specific percentages of client connections to a particular host server. In addition, there are options available for routing all requests from a particular client to the host that handled that client's first request.

Host servers in the cluster exchange “heartbeat” messages so that they can maintain consistent information about what host servers are members of a cluster node. If a host server fails, client requests are then re-balanced across the remaining host servers. Each remaining host server will then handle a percentage of request proportional to the percentage which was specified in the initial configuration.

C. Dynamic Server Provisioning

Dynamic Server Provisioning inherits both the concepts of Adaptive Computing [5] and On-demand Computing [2]. The concepts rely heavily on the ability to increase and decrease service-oriented workloads based on actual business needs. When the business needs peak, our technicians must be able to dynamically increase the workloads they run to meet demand. When business needs are low, our technicians are then able to decrease the number of service-oriented workloads they run to minimize ongoing costs. In the world of x86, server virtualization gives our technicians the ability to run as many workloads as they require and to provision or de-provision workloads as needed based on actual business need.

Let us consider the example above:

Table I illustrated a typical Windows Application stream during a training session. This application is "parked" most of the time. Our lecturers may need to modify records and obtain data from the application, but this level of workload is very light in terms of the resources it requires. However, in a typical training session, we need to actually run the application and generate test records or files. When this process runs, the stream application requires more resources - many more resources.

<table>
<thead>
<tr>
<th>Resource</th>
<th>Sequence of Events</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low 1.</td>
<td>On Host Server 1, Windows™ Application is “Parked”</td>
</tr>
<tr>
<td>High 2.</td>
<td>On Host Server 1, Streaming Policy “Wakes” Application</td>
</tr>
<tr>
<td>Low 3.</td>
<td>Policy sees Host Server 1 will not have enough resources.</td>
</tr>
<tr>
<td>Low 4.</td>
<td>Policy finds other Host Server with available resources.</td>
</tr>
<tr>
<td>Low 5.</td>
<td>Policy moves workload onto Host Server 2</td>
</tr>
<tr>
<td>Low 6.</td>
<td>Windows™ Application runs on Host Server 2 as well as Host Server 1</td>
</tr>
<tr>
<td>Low 7.</td>
<td>Policy removes application back to Host Server 1, and “parks” Host Server 1 again.</td>
</tr>
</tbody>
</table>

In a virtualized world, the application on a host server is able to be “parked” during the time it requires few resources. Then, when it needs more resources to run, it can be moved dynamically to another host server and it can be made sure the application has the resources it requires for this job. Finally, when the job is complete, the application can be moved back to a parked state. For this process to work, our virtualized infrastructure is required to perform two tasks:

1) Move virtual machines (VM) from host to host without service interruption. Hence, using cluster nodes.

2) Support either the assignation of a variable amount of resources to a VM, or modify the resources assigned to a VM during its operation.

The above example highlights the fact that, with virtualization and proper capacity planning, academic systems are able to cope with sudden surge in online demands in any particular period of time.

D. The Presentation Layer

The presentation layer is a central location to all users either local or remote access from. These users are able to access; use and manage their virtual workloads, for example, via terminal service clients utilizing the Virtual Desktop Infrastructure (VDI). It is a server-centric computing model that borrows from the traditional thin-client model but is designed to bring system administrators and end users these benefits: (a) the ability to host and manage desktop virtual machines centrally in the data center whilst giving end users a complete PC
desktop experience. (b) Multiple options are available for centralized desktop and application deployments and each option having its own pros and cons.

Conventional session-based deployment option like terminal services is currently the most scalable and high performance option. It is based on the capability of Windows™ server to create multiple sessions on a single instance of OS. Administrators install and manage a full desktop on centralized servers in the data centre. Users connect to through virtualization technology. Restrictions can be applied by the administrator to a specific application (i.e. presentation) in place of a full desktop. This option will certainly benefit our students who require access to an entire desktop that contains fewer simple applications or specific applications which is not suitable for the client’s PC.

Our second option is VDI. Client OS is hosted on servers in the data centre. The user then connects to it remotely using presentation virtualization. User can be assigned a dedicated VM so that all user data (e.g. My Documents) and profile information, hence, personalization, are retained on the VM image. This deployment is suitable for knowledge workers (e.g. lecturers, software developers or testers) who require administrator rights to have full control over their virtual desktop to deploy their own applications and to customize environment.

A third option is through pooled VMs that are identically configured. Pooled Virtual Desktops (PVD) are best suited for students (task workers) who need to work on some standard applications and do not require personalized desktop configuration or customization. In this configuration, when a user’s session ends, the data is not stored on the virtual machine. In a typical configuration, folder redirection is used to save data to another server so it is available when the user logs in next time but no configuration data is saved between sessions. This VM based option eliminates the limitations of session-based deployments of administrative rights and application compatibility. However, it is less efficient in terms of the number of users that can be accommodated on the server hardware. For instance, session-based infrastructure can accommodate almost ten times the number of users than a VM would on the same server hardware for the students’ scenario.

At this planning stage, our team is yet to choose a single and most appropriate presentation technology. However, due to our complex requirements, it is likely that we will implement a mixture of presentation layers for our own purpose.

IV. OUR PROPOSED INFRASTRUCTURE DESIGN (SIMPLIFIED VERSION)

Based on what have been discussed above a simplified version of our infrastructure design has been proposed. Fig. 4 illustrates our virtualized private cloud our project team wishes to implement and we have placed great emphasis on application streaming, scalability and clustering.
In Fig. 4, we included four key cluster nodes: web applications, SQL data access, streaming server and terminal service servers. Each cluster nodes function independently and is managed by an operation called Server Provisioning by utilising the cluster’s heartbeat listening device. Another function for Server Provisioning is to load or remove the application image which is stored in the file image library. When a sudden surge in demand of a server resource is encountered, Server Provisioning will inform VMM console to load an image file onto a spare server before it is moved into a cluster node, or vice versa. The parameter settings for the thresholds are configurable.

A new Windows™ application, waiting to be virtualized, can first be sequenced and packaged before it is being stored in the Image File Library. Storage Attached Network (SAN) is used as our preferred virtualized storage solution. Client PCs or laptops can access the network either locally or via the Internet. Let us assume that we will be using session based terminal service clients to access our network and application streaming facilities.

A. The Potential Pitfalls
There are a number of potential pitfalls our project team is required to resolve. First, it is the Hypervisor. Our team intends to look into dynamic memory management within virtual machines, hot-adding of resources to virtual machines and more.

Second, it is the sequencer. Despite the sequencer being powerful there are outstanding questions we must bear in mind before pursuing application sequencing and packaging. These questions are:

1) What application components are needed and will be installed?
2) What updates, such as adding new files to the package, need to be performed in the sequencer after the installation?
3) What post-installation configuration steps need to take place in the sequencer?
4) What do users commonly do with this application immediately after its launch?
5) What file types are associated with this application?
6) And finally, does this application do something that the sequencer currently does not support?

Third, there are limitations in session based presentation layer deployment. Users cannot get administrative rights to the machine as that can impact users working on other sessions on the same machines. There are also application compatibility issues due to OS compatibility with the applications and application concurrent execution.

Fourth, our project team will have to evaluate the options and then decide what works best for our business needs. We also noted that centralized desktop is not suitable in all circumstances. For example, if the desktop is primarily used for high-end multimedia applications, rich desktop is still the best choice. Instead our project team will use a mixture of rich and centralized desktop and bring in the benefits of both options.

V. OUR PRIVATE CLOUD
This section concerns the wider usages of our Private Cloud. We attempt to evaluate our Cloud in the context of classroom and learning, administration and operations, and lastly applications development.

It has been proved [7] that, with cloud computing, academic institutions can open their technology infrastructures to businesses and industries for research advancements. The efficiencies of cloud computing can help academic institution keeps pace with fast growing resource requirements and energy costs. Private Cloud can also enable academic institution to teach students in new, different ways and help them manage projects and massive workloads.

A. Classroom and Training
It has also been claimed [10] that, with Cloud Computing, educators are able to respond dynamically to the evolving requirements of their communities. Students are able to connect their personal mobile devices to campus services for learning. Teachers and faculty are also able to have greater access and flexibility when integrating technology into their classes. Researchers are able to have instant access to high performance computing systems, without the responsibility of managing a large server and storage farm.

B. Administration and Operations
In the context of administration and operations, it has been acknowledged [11] that, with Cloud Computing, IT staffs in universities and colleges are finding it easier to maintain desktop computers, laptops and various other devices and infrastructures. Increasing tight budgets limit what education institutions can spend on computers for students, teachers and administrators. Constantly evolving technology and the periodic nature of funding can result in a lack of standardization across a district, campus or system but all these can be mitigated with the help from Cloud Computing. Also, the geographic distribution of schools or campuses no longer stretches the ability of technical staff to give hands-on support as they have to travel to make even simple fixes in PCs – the configuration of which they don’t know until they arrive on site. At the same time that computers have proliferated, so have applications and the devices storing the data of the institution and individuals. Managing applications and data is more complicated than ever before but with the help from Cloud Computing all these issues are able to be resolved.

C. Applications Development
For applications development in Cloud Computing, scalability and ease of maintenance are keys to success. Our project team has devised a 7 tiers approach to software development as illustrated in Fig. 5. It is worth noting that Tier 4 is a conventional middleware or computation layer which will be used for database access, on-the-fly computation as well as tasks queuing. For
example, workflows can be built in the Intranet while business logics including computation can be built into the middleware.

**Figure 5. Our private cloud and its software development tiers**

<table>
<thead>
<tr>
<th>Tier 7</th>
<th>User Interface (OS + Web)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tier 6</td>
<td>Web Applications</td>
</tr>
<tr>
<td>Tier 5</td>
<td>Web Server</td>
</tr>
<tr>
<td>Tier 4</td>
<td>Middleware or Computation</td>
</tr>
<tr>
<td>Tier 3</td>
<td>Database</td>
</tr>
<tr>
<td>Tier 2</td>
<td>Storage</td>
</tr>
<tr>
<td>Tier 1</td>
<td>Operating System (OS)</td>
</tr>
</tbody>
</table>

VI. CONCLUSIONS

Virtualization for academic computing delivers instructional applications and data to students and faculties on any computer over any network connection at the lowest possible cost. There are a range of benefits for virtualization and they are:

1) **Reduce costs and significantly cut down on the amount of I.T. management hours required to manage desktops and applications.**

2) **Quickly provision, manage and deliver desktops with a better student and lecturer end-user experience and the lowest cost of ownership through simplified life cycle.**

3) **Securely deliver course and module based access that can be managed at the server level.**

4) **Eliminate dependence on the traditional computer laboratories.**

Virtualization and Private Cloud Computing rely on implementation of all layers as clearly defined in the Virtualization Model before a Cloud infrastructure is formed. Part of our solutions involves application streaming by utilizing sequencing and packaging where Windows™ applications are streamed in small packets on-demand at real time. Our solution also mention network load balance and clustering technology where managing cluster nodes can be made by utilizing heartbeats listening. Dynamic server provisioning is also used in our solution which is based on real-time demand on the server load. The presentation layer offers centralized desktop management for all users with the aim to offer secure and quick provision to access system, applications and files. Cloud computing is an architectural approach to providing services across a web-based infrastructure. It holds the potential of providing a dynamic infrastructure that can respond to as demands increase. This approach enables the sharing of technology resources in our private model. Cloud computing creates a virtualized infrastructure which can provide a superior user experience, and is characterized by new, Internet-driven economics. Finally, based on virtualization and cloud computing technology currently on offer, our project team is expected to use a mixture of solutions and bring in the benefits of all options available.

ACKNOWLEDGMENT

The authors wish to thank their colleagues and external advisors for their help and support, in particular, Ryan Tunstall (Global I.T. Director, HIT Entertainment Limited, UK), James Herbert (Head of I.T. Infrastructure, HIT Entertainment Limited, UK), Katie Forde (I.T. Director, Guinness World Records Limited, UK), and all I.T. staffs at Zhong Shan Polytechnic.

This paper is supported by National 973 (2007CB310804), National Natural Science Fund of China (60743001), Best National Thesis Fund (2005047), Natural Science Fund of Hubei Province (CDB132), "Eleventh Five-Year" Adult Education Research Plan of Guangdong Province of China (yxj092015), "Eleventh Five-Year" Education Research Plan of Guangdong Province (2009jk002) and 2010 Education Scientific Research Plan of Zhongshan City (D10143).

REFERENCES


DENG Jiabin is a lecturer at the Computer Engineering Department of Zhong Shan Polytechnic. He received a Bachelor Degree in Computer Engineering from Hubei University in 2005, then a Master Degree in Computer Engineering from Wuhan University in 2007. His research interests include artificial intelligence, data mining, and complex network. His research papers are also indexed by IEEE CS, including "An improved classification algorithm on teaching evaluation", "An Improved Algorithm of Bi-cubic Triangular Curved Surface-splicing for Curved Surface Reconstruction Based on Hermite Surface", and "A New Probe Into The Colour Modeling Of Colour Output Devices Based On The Method Of Five-Level Non-Uniform Partition And Pyramid Interpolation".

Juanli Hu is currently lecturing at the Computer Engineering Department of Zhong Shan Polytechnic. She has been working as Director of Teaching ( Fundamental Computing Research) since 2006. She received her Bachelor Degree in Computer Engineering from Xi'an University of Technology in 2000, followed by a Master Degree in Computer Engineering from Xi'an University of Technology in 2005. Her research interests include signal, information processing and data mining. Some of her research papers are indexed by IEEE CS, including "An Improved Algorithm of Bi-cubic Triangular Curved Surface-splicing for Curved Surface Reconstruction Based on Hermite Surface" and "A New Probe Into The Colour Modeling Of Colour Output Devices Based On The Method Of Five-Level Non-Uniform Partition And Pyramid Interpolation".

LIU Anthony Chak Ming is lecturing at the Computer Engineering Department of Zhong Shan Polytechnic, PRC. Since graduated at Cardiff University and Open University in 1989 and 1992 respectively Anthony had worked in the field of software application engineering in the United Kingdom. In excess of 100 various I.T. projects Anthony had worked on for some well-known multi-national companies he was responsible for providing a wide range of technical services and I.T. solutions. After spending 28 years abroad, in March 2009, Anthony returned to China permanently. His recent projects include Virtualization, Hosting Solutions for Data Centre, Software Application Design and Optimization, SQL Database Design and Optimization. His most recent research interest is Private Cloud Computing for large corporate.