vmware®

VMware ESX Server 3.0

Improving Scalability for Citrix Presentation Server

Citrix Presentation Server administrators have often opted for many small servers (with one or two CPUs) to run Citrix Presentation Server server farms. Due to the limitations in the underlying operating systems, it has been difficult to scale the number of Citrix Presentation Server clients that can be supported on a server as the number of CPUs in the server is increased.

With multiple-core processors becoming the norm, and servers with four to eight CPUs providing the best price point for servers, it is important for applications in the data center to be effective in scaling to servers with larger numbers of CPUs. Using VMware ESX Server it is possible to use larger servers as the infrastructure for a Citrix Presentation Server implementation. The purpose of this paper is to characterize the performance of Citrix Presentation Server running inside multiple VMware virtual machines on an ESX Server 3.0 host. It covers the following topics:

Executive Summary

Administrators responsible for Citrix Presentation Server configuration are tasked with using their hardware resources efficiently. A useful guideline for assessing hardware capacity is the number of users who can be supported with 80 percent CPU utilization on the host system. The primary aim of our tests was to evaluate the proposition that virtual machines running under VMware ESX Server 3.0 could improve resource utilization on modern servers equipped with eight CPU cores or more — an increasingly common configuration in organizations using Citrix Presentation Server. Our tests demonstrated that the total number of users supported on an eight-core configuration based on ESX Server 3.0 scales in linear fashion as virtual machines are added.

vmware®

Test Setup

The tests used the following hardware and software configuration:

Server Hardware

- Host computer: Hewlett-Packard DL 585
- CPUs: 4 AMD Opteron 875 dual-core processors at 2.2GHz (total of 8 CPU cores)
- RAM: 32GB
- Hard drives: 4 150GB SCSI 15,000 RPM drives with a Hewlett-Packard Smart Array 5i Plus controller.
- Network: 1Gbps Ethernet adapter

The server and client systems were connected to the same network switch.

Server Software

- VMware ESX Server 3.0

Server Virtual Machine Configuration

- CPU: 1 virtual CPU
- RAM: 3.5GB
- Operating system: Windows Server 2003 Enterprise Edition (32-bit)
- Citrix software: Citrix Presentation Server Version 4

Client Hardware

- Client computers: 2 Dell 1850 and 2 HP DL385 systems
- RAM: 4GB in each system
- Network: 1Gbps Ethernet adapter
- Load: 2 virtual machines are driven by each client system

The server and client systems were connected to the same network switch.

Client Software

- Operating system: Windows Server 2003 Enterprise Edition (32-bit)

vmware®

Test Suite

The test used was the Citrix Server Test Kit, a user simulation infrastructure provided by Citrix that offers settings to run various applications. For these tests we used standard users, as defined in the test kit, running Microsoft Word.

Standard users start Microsoft Word and enter one of two documents. Auto-correction and spell checker options are turned on, as would be the case for a normal user. Under normal circumstances the user types up the document in 11–15 minutes. After a very short pause, the user starts the task again. The test is set up by default to run for 80 iterations; for our purposes the test was stopped after a steady-state period of 30 minutes had elapsed after the last user had logged in.

Benchmark: Simulated Users Compared to Actual Users

In these tests the users are simulated through software. The users thus tend to be very uniform in their behavior, keystrokes from all users arrive at regular intervals, all users are uniformly active throughout the test, and a single application is used.

In the real world users tend be less uniform in their activity level. There are often periods when activity levels are relatively low, while a user is engaged in a conversation, is reading a document online or offline, or has stepped away to a meeting. On the other hand, real users' work patterns also tend to include flurries of activity when they use many applications at the same time, open and close documents in rapid succession, cut and paste from one application to another, and otherwise generate higher loads than a simulated user does.

Because of these differences, the number of users supported in the real world per CPU would usually be lower than the number presented in this paper. This ratio of simulated users to real users depends on the nature of use and activity level of the real users.

Testing Criteria

For the sake of comparison, we wanted to see how many users could run on a given configuration and what the average CPU utilization was when these users were active. Based on published results from benchmark tests from Citrix and experience with user failures, we determined that running at or around 80 percent CPU utilization was ideal for the purpose of these measurements. We determined through a process of trial and error that on ESX Server 3.0 running 140 users on a virtual machine with one virtual CPU and 3.5GB of RAM would cause utilization to reach 80 percent on our test platform.

Test Ramp Up

We found that the process of logging users on had a significant impact on CPU utilization. In these tests it is therefore important to have the users log on in a regulated manner. The ramp-up schedule used for these tests was as follows:

Users Logging On	Delay between Logons
User 1 to user 5	30 seconds
User 6 to user 50	60 seconds
User 51 and higher	120 seconds

Failure Detection and Recovery

As CPU utilization on the server increases, the time it takes the server to complete expected actions falls beyond the limits set up in the test kit. For example, a heavily loaded server may not open the Word application in time for it to open the subsequent document. The driver program

is a load generator by design and continues to send keystrokes assuming the application is in the state it expects it to be. However, because the keystrokes are not being recorded in the application they are supposed to go to, this user is no longer contributing load as expected and is considered to have failed.

One of the ways to detect failure is to look at the desktop for each running user to make sure there is still activity happening on that desktop. This proved tedious, especially for the larger tests when we were running many hundreds of users. As an alternate method, we parsed the output logs from the all the users. These logs record the start and end of each iteration of the test task. Users were considered to be running a valid iteration if there was a record of the iteration ending. All iterations that did not have a logged end time were discarded as failures.

The other type of failure often seen in these tests is one that causes a user to consume CPU cycles at a very high rate. If this situation occurred, the user consuming excessive CPU resources was terminated and another user added at the end of the test to bring the number of running users to the intended level. The failure rate tolerated in the tests was on the order of 1–2 percent. In other words, if in a test we planned to start 100 users and more than two users failed and had to be terminated, we discarded the test and re-examined the test configuration to see if we could make changes to help the test succeed. We then attempted the test again. The results we present here are a result of this iterative process to determine what volumes of users could run successfully.

In the output logs, every start of iteration signifies an active user and every end of iteration indicates a user that has become inactive. There is a small pause between the end of one iteration and the start of the next one. As a result of this logging methodology, the number of active users shown in the charts can be slightly lower than the number of users simulated in the test.

vmware®

Test Results

Running Citrix Presentation Server on a virtual machine has some overhead, especially in memory management, due to the nature of the application. ESX Server 3.0 shows considerable improvement in minimizing this overhead when compared to ESX Server 2.5, and VMware continues to investigate opportunities for further improvement. Figure 1 shows results for a single virtual machine running on both ESX Server 2.5 and ESX Server 3.0 hosts. It shows virtual CPU utilization and the number of users running. These tests were run in identical virtual machines with one virtual CPU and 3.5GB of memory.

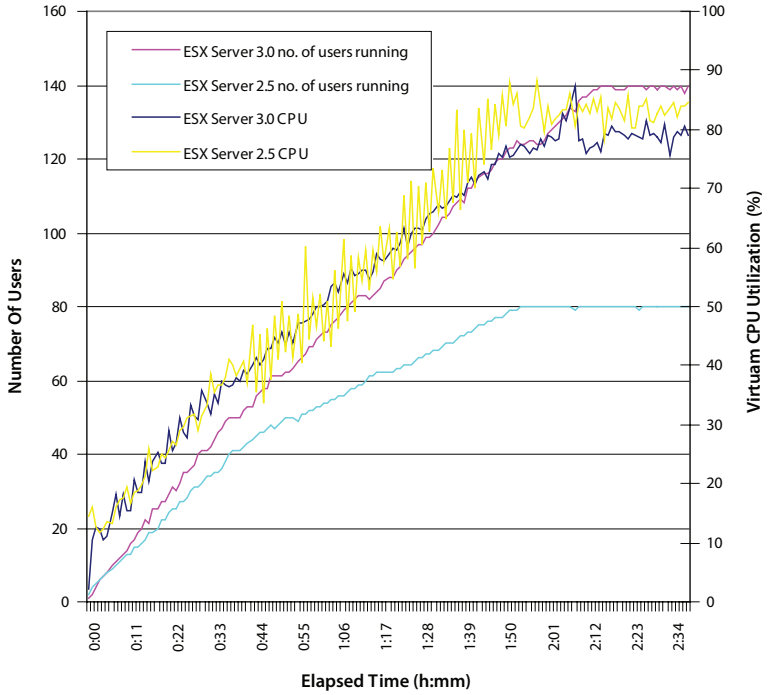

Figure 1: ESX Server 2.5 compared to ESX Server 3.0

As can be seen above, for ESX Server 2.5 we are able to run the test with 80 users with virtual CPU utilization at 85 percent. When the same virtual machine is run under ESX Server 3.0, the test runs successfully with 140 users at 80 percent CPU utilization for the virtual machine.

Scalability with Multiple Virtual Machines

The benefits of running Citrix Presentation Server in ESX Server virtual machines become readily apparent when the test is run with multiple virtual machines. We set up eight identical virtual machines. Figure 2 shows the results for runs with one, two, four, six, and eight virtual machines. As shown in the figure, the number of users scales linearly as virtual machines are added.

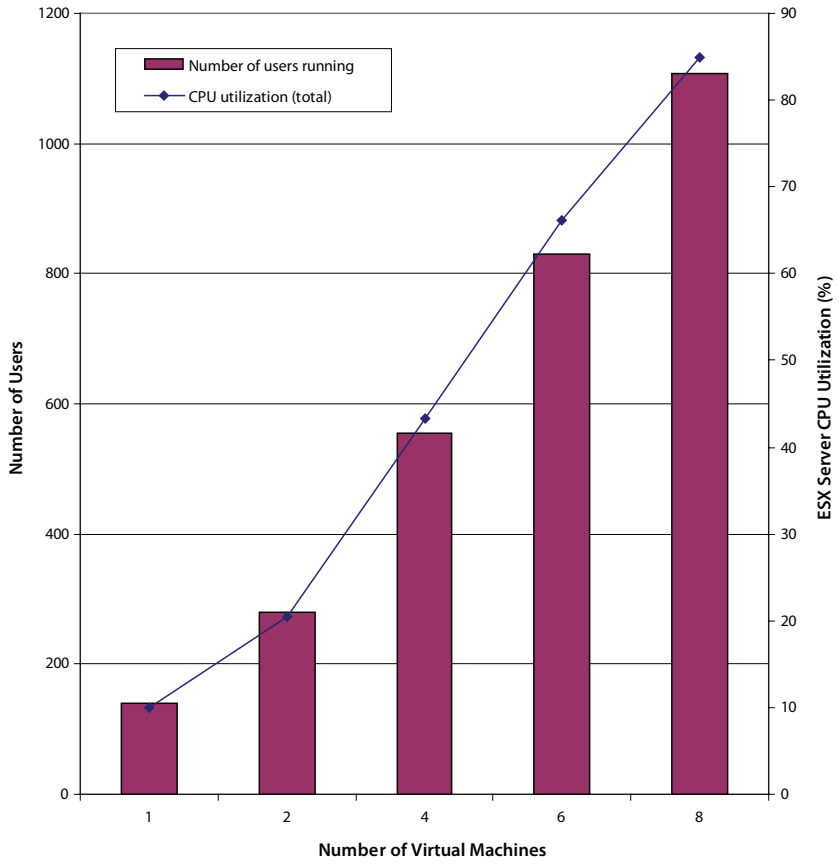

Figure 2: ESX Server 3.0 scalability

For the configuration with eight virtual machines, the run was successful only after each individual virtual machine was pinned to run on a specific CPU.

Conclusions

The results of our tests clearly demonstrate that ESX Server 3.0 provides an excellent platform for customers looking to deploy large installations of Citrix Presentation Server using commodity server hardware and fully exploiting the benefits from multiple-core CPUs. The underlying operating system limits placed on Citrix Presentation Server can be avoided by running multiple copies of the operating system on one host using VMware virtualization technology, allowing for more effective utilization of these systems This approach allows datacenters to use more compact and power-efficient systems to deploy Citrix Presentation Server.

The number of users supported in the real world will depend on the applications used and the activity levels of users. The numbers provided here demonstrate scaling and provide a comparison across ESX Server versions — ESX Server 2.5, ESX Server 3.0, and multiple virtual machines running under ESX Server 3.0 — and do not give absolute numbers for user loads that can be supported on a CPU.

VMware, Inc. 3145 Porter Drive Palo Alto, CA 94304 www.vmware.com

Revision 20061026 Item: ESX-ENG-Q406-335