Comparison of virtualizing Citrix XenApp 5.0 with XenServer 5.5 on Intel Xeon X7350 & X5570 host servers

www.citrix.com

Introduction

By far the largest segment of Citrix customers are those who use Citrix® XenApp™. The largest segment of the XenApp community are those using 32-bit XenApp applications. While 32-bit XenApp has been known for many years for reliability and performance, it's also known for being somewhat limited in scalability due to the fact that it relies on a 32-bit Windows Server® operating systems (OSs), based on the fact that the 32-bit Windows® OS kernel can only support a relatively small number of server resources. Depending on the application used, it's fairly common for XenApp servers to have only 2 CPU cores with 4GB RAM. Adding more resources often doesn't produce better results (measured in terms of the number of concurrent XenApp users per server). This sometimes presents a challenge to the XenApp administrator whose goal is to add more users while also having to reduce power consumption and server footprint in the datacenter. Without upgrading applications to 64-bit XenApp, which can be time consuming, the simplest answer to this problem is to virtualize the server hardware. By running multiple, small 32-bit XenApp servers hosted on today's large, powerful 64-bit server systems, administrators are able to achieve remarkable levels of server consolidation, with little or no impact to user performance. Coupled with the inherent manageability and flexibility benefits of virtualization, it's no surprise that increasingly large numbers of XenApp farms are going the way of server virtualization—with most of them doing so using Citrix® XenServer™.

Earlier this year, as Intel® released its new Xeon® 5500 (code named Nehalem) series processors, we demonstrated some remarkable test results showing a solid 53 percent performance improvement between 5400 and 5500 based servers when running a DBHammer SQL Server 2008 workload http://community.citrix.com/pages/viewpage.action?pageId=65732872. It stood to reason that we might find the same kind of performance enhancement with a XenApp workload. Customers seeking to virtualize their XenApp farms have a wide choice of hardware platforms when using XenServer to host XenApp virtual machines (VMs). We created a test project to validate our assumption that the Intel Xeon 5500 would be the logical choice for the largest density of XenApp users in the smallest possible space. For comparison purposes, we chose to benchmark the performance of a physical XenApp server using an earlier generation Intel CPU server, the Xeon X7350. We selected the Xeon X7350 since, like the X5570, it too is a quad core CPU. In this case, both ran at the same clock speed with the same relative power consumption. Could we successfully virtualize these 32-bit XenApp VMs on dual quad core X7300-based hosts instead of the latest generation Xeon X5500s? Absolutely. However, it was our belief that we would see the kinds of performance gains hosting XenApp VMs that we previously saw in our Xeon X5500 test running SQL Server® 2008. Assuming we did see a substantial improvement, the resulting server consolidation might be that much better as well.

Project Goals

The goals of this test project were to:

- Determine the maximum number of concurrent 32-bit XenApp users on an X7350-based physical server using a standard 2 core, 4GB server configuration.
- Determine the maximum number of concurrent users on a dual quad core, 16GB Xeon X7350 server if it were used as a XenServer host to support multiple 32-bit XenApp VMs with the same configuration.
- Determine the user density results for a single VM and the maximum number of VMs possible when using a dual quad core Xeon X5570-based XenServer host as compared to the Xeon X7350 host.

We also had an additional goal of understanding the impact of hyperthreading on user density. Intel describes hyperthreading as "delivering thread-level parallelism on each processor resulting in more efficient use of processor resources, higher processing throughput and improved performance." We found hyperthreading to be a positive benefit in the earlier SQL Server performance tests and so we wanted to see if we would see similar benefit with a XenApp workload. Would hyperthreading allow us to run twice as many VMs on a single host? If so, how would the density of each VM be affected? To make this determination, we compared performance both with and without hyperthreading.

Overall, we expected not only to demonstrate that virtualizing 32-bit XenApp farms using Xeon X7350 host servers would create a substantial server consolidation opportunity, but that doing the same on a Xeon X5570 would create even greater server consolidation.

Test Environment

1. Physical Citrix XenApp 5.0 Server: Single XenApp server, 2 CPUs, 4GB RAM.

   Hardware:

   - Dell® PowerEdge® r900, Intel Xeon X7350, 2.93GHz quad core CPU (two CPU cores used), 4GB DDR2 RAM (set by boot.ini)

   Software:

   - Microsoft Server 2003, Enterprise Ed., SP2, 32-bit
   - XenApp 5.0

   Storage:

   - NetApp™ FAS3140, configured for iSCSI

2. Citrix XenServer 5.5 virtual host server. Eight virtual Citrix XenApp 5.0 servers on a single virtual XenApp server with 2 vCPU cores & 3.6GB vRAM

   Hardware:

   - Intel Tylersburg-class whitebox server, Intel Xeon X5570 @ 2.93GHz dual quad core CPU, 32GB DDR3 RAM

   Software:

   - XenServer host: XenServer Version 5.5.
   - Microsoft Server 2003, Enterprise Ed., SP2, 32-bit
   - XenApp 5.0

   Storage:

   - NetApp FAS3140, configured for iSCSI

3. Citrix® EdgeSight® for Load Testing Servers

   Hardware:

   - XenServer host for ESLT launcher VMs: Intel whitebox server w/ four Intel Xeon E7450 @ 2.4GHz hex core CPUs, 32GB DDR2 RAM. Eight virtual ESLT launchers with 2vCPUs and 3.6GB RAM per VM (one per XenApp server, physical or VM), Windows XP, x32, SP3.
   - XenServer host for ESLT controller VM: Intel S5000PAL whitebox server w/ Intel Xeon E5345 2.33GHz quad core CPUs, 8GB RAM. One controller required for each test configuration.

   Software:

   - EdgeSight for Load Testing ver 3.5
   - Microsoft® Office 2007, Excel® and Word

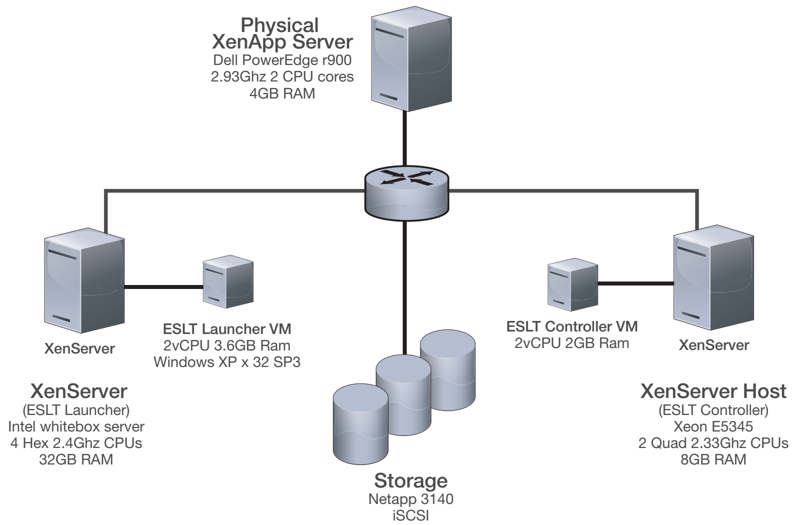

Figure 1. Physical XenApp server configuration

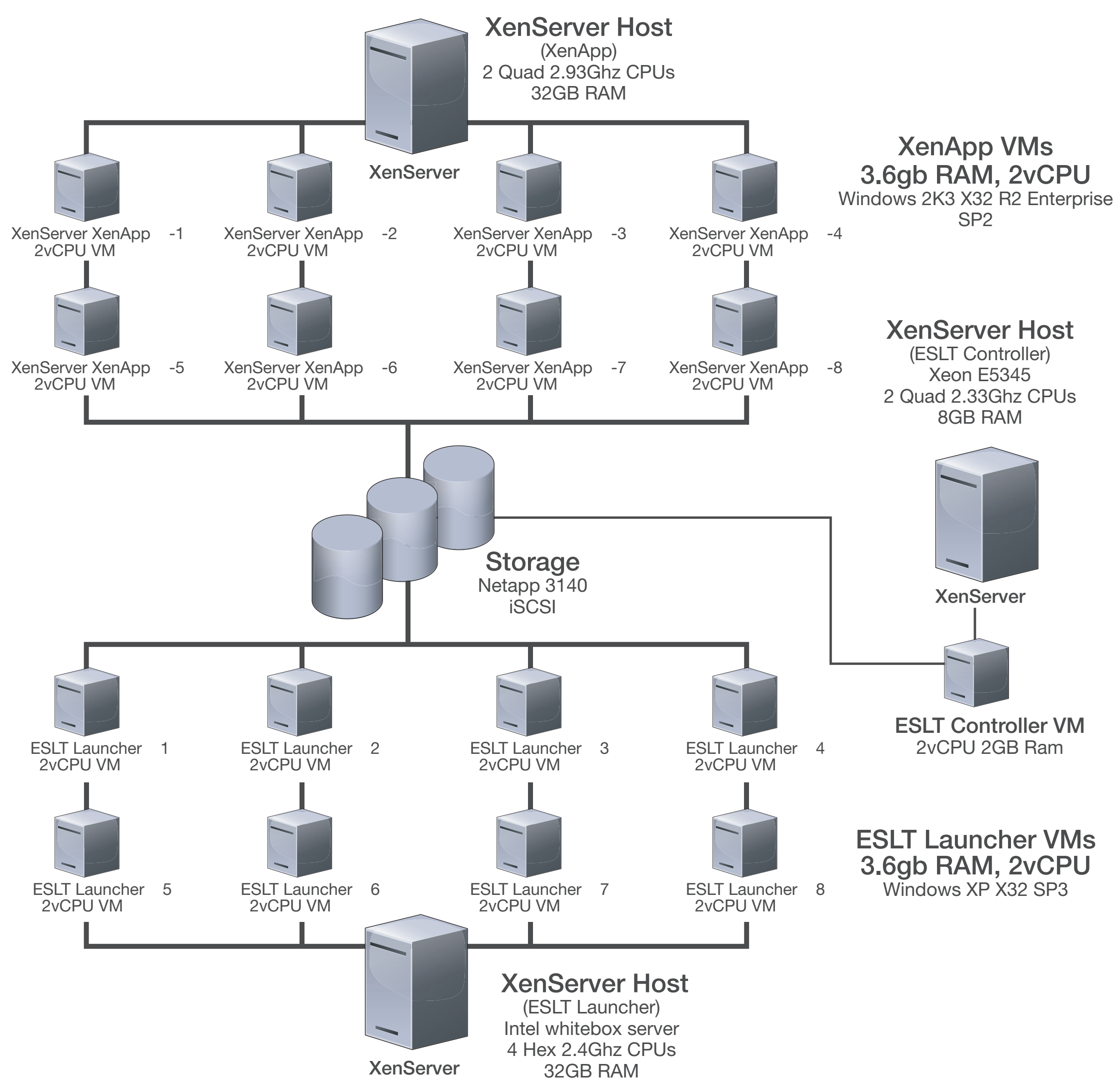

Figure 2. Virtual XenApp server configuration

Methodology

We began by building a physical 32-bit XenApp server on the X7350-based server. Because of the limits of the 32-bit OS, we configured it to use only two of the CPU cores and 4GB of RAM. Although the server itself had far more CPU cores and RAM available, 32-bit XenApp wouldn't effectively be able to use it. This XenApp configuration became the standard by which we would build all of the 32-bit XenApp VMs to be tested (although each VM would actually only have 3.6GB of RAM to work with accounting for the need for RAM to support virtualization overhead). We would measure the number of concurrent XenApp users on the physical XenApp X7350-based server and then calculate the maximum number of users we could run on a dual quad core, X7350 and X5570 XenServer hosts.

In the XenServer setup screen, we selected the option of running XenApp which automatically configured the VMs with the appropriate amount of shadow memory for XenApp workloads.

For load testing, we used Citrix EdgeSight for Load Testing (ESLT) version 3.5. ESLT is an automated load and performance testing solution developed specifically for generating test workloads in XenApp environments. ESLT generates virtual user ICA sessions on the target XenApp servers. The number of launchers required will vary based on the target virtual user load. ESLT launchers report session information back to the controller for run-time and post run-time analysis. ESLT enables administrators to predict how systems will cope with high levels of user load. By simulating hundreds of virtual Citrix users and monitoring the responsiveness of the system under test, it allows the administrator to determine how the current configuration and hardware infrastructure will support anticipated demand.

The test workload for this project was Microsoft Office 2007 (MS Word and Excel). Tests were performed by using EdgeSight for Load Testing to generate a steady user load inside the MS Office Suite, and then measuring the amount of time it took to perform a sequence of (automated) actions. We ran our measurement script first as a single user to generate a baseline for performance. We then determined the point at which multiuser performance became unacceptable and we declared the maximum number of sessions to be reached. Load was generated in increments of 4 on the physical XenApp server and 16 on the multiple XenApp VM server (est. 1.5 user logins per min.), applying rate control to our user logins to avoid flooding the server.

We established a baseline of 25 seconds for users to login, run a standard MS Office task script, and then logout (including network connect time). We added users until the threshold to run this sequence reached a latency of 30 percent, at which point the server was deemed to be at capacity.

Results

X7350 physical XenApp server

The results for the physical XenApp server were not all that surprising. 47 users on a 2 core, 4GB server is respectable and fairly common for XenApp users with this kind of workload. We found that it suited the purpose of this test study rather well as a result, since many XenApp administrators would likely see themselves as having a similar level of user density.

Results averaged over three consecutive test runs of 120 user connections.

- ESLT script runtime = 25 seconds
- Target threshold = 30% latency level (32.5 seconds)
- 47 simultaneous XenApp client sessions achieved before the 30% latency level was reached
- Average CPU usage = 39.2%
- Maximum CPU usage = 72.3%
- Maximum RAM usage = 2.19GB of available 4GB

We determined that the difference between a physical and virtual XenApp server running on the X7350 would be minimal, due to the fact that the limit to the number of users was driven more by the 32-bit Windows OS kernel than by server CPU or RAM. Assuming that a customer were to scale out a X7350 using virtualization, then a dual quad core, 16GB server would support four VMs, each with 2 cores and 3.6GB RAM, producing a total of 188 simultaneous users (47 x 4). Therefore, virtualization in this case would offer the potential for a 4:1 server consolidation.

Virtual XenApp server – X5570 host, with hyperthreading

We ran our first virtualized XenApp test with hyperthreading activated and then repeated the test again with it turned off. With hyperthreading, the first thing we noticed was that even though there were only 8 CPU cores on the X5570 host server, XenServer was able to see 16 vCPU cores as resources available to be assigned to VMs. As a result, we were able to successfully run a maximum of eight VMs, each with the necessary 2 vCPU cores, generating an average of 69.25 users per VM for a total of 554 users. If migrating from 32-bit physical servers based on the X7350, then a server consolidation ratio of 11.8:1 would have been achieved.

When compared directly to the performance of the X7350, we saw that average CPU utilization was slightly higher on the X5570, while maximum CPU utilization was slightly lower. We attributed this to the fact that hyperthreading causes physical CPUs to work harder in order to produce twice the number of logical CPU threads as without hyperthreading. Intel is able to balance processing using both physical and logical CPU threads resulting in higher average CPU utilization while being better able to minimize overall peak utilization. In addition to the obvious ability to now run more virtual systems with these additional logical CPUs, a second benefit seems to be that overall CPU predictability and stability is likewise achieved.

Results averaged over three consecutive test runs of 720 user connections.

- ESLT script runtime = 25 seconds
- 69 simultaneous XenApp client sessions on a single VM
- 554 simultaneous XenApp client sessions on 8 VMs of 2 cores, 3.6GB RAM each
- Maximum latency reached was 31.27ms (below 30% latency threshold)
- Average CPU usage was 44.02%
- Maximum CPU usage was 48.93%
- Average RAM usage was 2.48GB of available 3.6GB

Virtual XenApp server – X5570 host, without hyperthreading

When we ran the second virtualization test, this time with hyperthreading turned off, the number of users per VM increased slightly to 88. However, the maximum number of VMs was now only four, due to the fact that we now only had 8 vCPU cores to work with. As a result, the total number of users for the host was now only 352. If migrating from 32-bit physical servers based on the X7350, then a server consolidation ratio of 7.5:1 would have been achieved.

Results averaged over three consecutive test runs of 360 user connections.

- ESLT script runtime = 25 seconds
- 88 simultaneous XenApp client sessions on a single VM
- 352 simultaneous XenApp client sessions on 4 VMs of 2 cores, 3.6GB RAM each
- Maximum latency reached was 32.2ms (below 30% latency threshold)
- Average CPU usage was 36.0%
- Maximum CPU usage was 40.9%
- Average RAM usage was 1.55GB of available 3.6GB

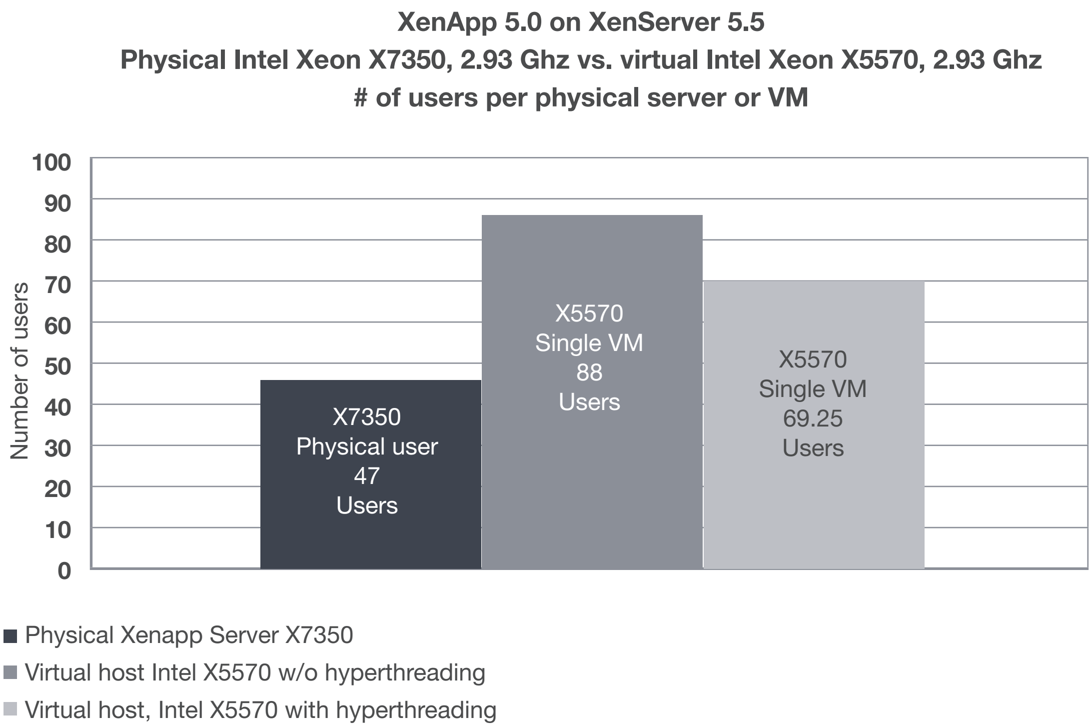

Figure 3. X7350 vs. X5570 single server user density

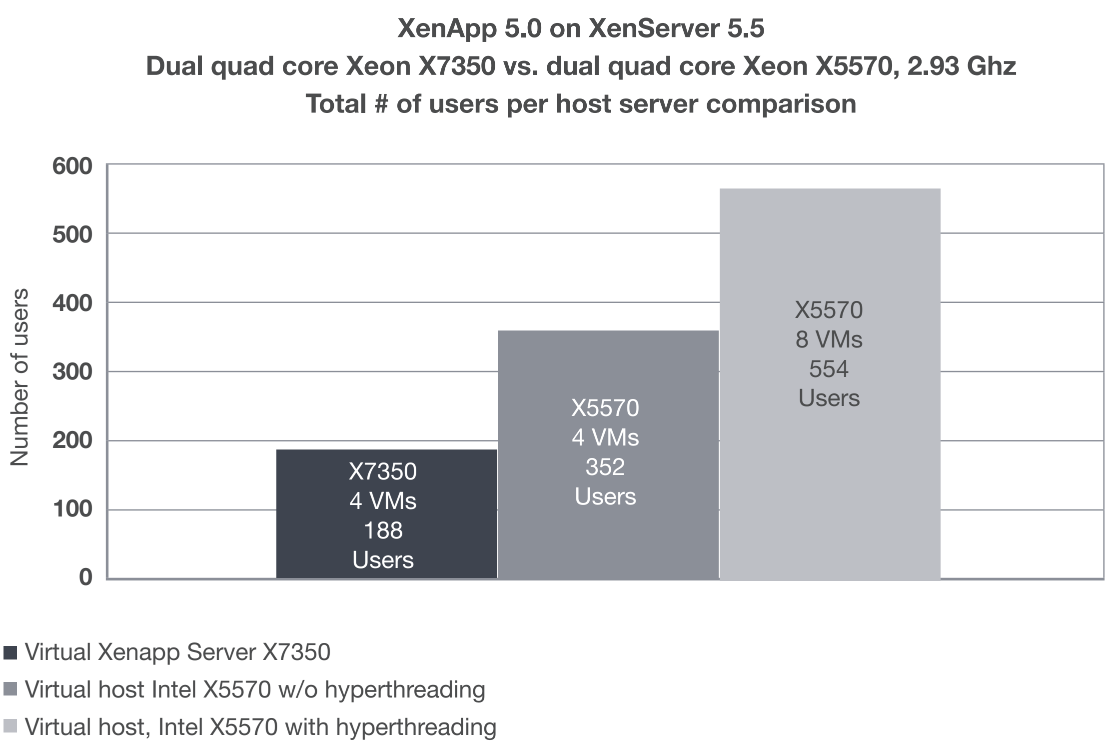

Figure 4. X7350 vs. X5570 maximum user density

Summary

In the end, we discovered that while hyperthreading doubled the number of assignable vCPU resources, it didn't directly translate to a 2:1 increase in the number of users per VM. That's a reasonable trade-off, since hyperthreading effectively doubled the number of VMs that we could create with the same number of CPU cores. Using XenServer 5.5, we were able to support 87 percent more users on the Xeon X5570 host server without hyperthreading as compared to a Xeon X7350 host server. However, as soon as we took advantage of hyperthreading, the number of concurrent users increased by nearly 200 percent over the X7350.

Users per server physical or virtual	Maximum # users per dual quad core XenServer host	32-bit XenApp server consolidation
X7350 = 47	X7350 = 188	X7350 = 4:1
X5570 w/o HT = 88	X5570 w/out HT = 352	X5570 w/o HT = 7.5:1
X5570 with HT = 69	X5570 with HT = 554	X5570 with HT = 11.8:1

While 32-bit XenApp will virtualize well with XenServer using Xeon X7350 host servers, it's clear that the ability of the X5570 to support more users per VM and, with hyperthreading, support twice the number of VMs thus producing far greater user density on a single host, is a decided advantage. If the XenApp administrator's goal for virtualization is server consolidation, reduction in power usage and server footprint, then it's hard to argue why one wouldn't consider the Intel Xeon X5500 as the ideal host server to do it with.

As we've seen here, the promise of Intel's Nehalem technology is being realized in some very practical ways. As a result, the performance bar for 32-bit XenApp, when virtualized with XenServer, is now higher than ever. The opportunity for 32-bit XenApp server consolidation has likewise never been better.

0809/PDF