

Server and Storage Sizing For VMware VDI: A Prescriptive Approach



Table of Contents

About this Document	3
Introduction.....	3
Baseline Existing Desktop Environment.....	3
Performance Monitoring Tools.....	4
Key Attributes to Monitor.....	4
Profiling the Target Desktop.....	4
Estimate Hardware Needs.....	5
Estimating CPU	5
Estimating Memory	6
Storage – Performance.....	7
Storage – Capacity	8
Build Proof-of-Concept Infrastructure	10
Lab-based Case Study Test Environment.....	12
Validate Hardware Estimates	12
CPU Observations.....	13
Memory Observations.....	14
Storage Observations.....	16
Impact on End-User Experience	17
Conclusion	19
Other Resources.....	19

About this Document

This paper is designed to help organizations size their server and storage resources for a Virtual Desktop Infrastructure (VDI) implementation. Previous papers on this subject have documented attempts to produce a single “rule of thumb” estimate that would hopefully be applicable across a broad spectrum of customers and environments.

This document approaches the challenge of resource sizing from a different perspective: rather than supplying a rough estimate of how many virtual machines can be housed on a given server, this white paper will lay out a methodology by which IT administrators and architects can reach an accurate estimate based on the unique requirements of their installation.

Note: It is assumed the reader is already familiar with VDI and the related benefits of the technology. If an introduction to VDI is necessary, please see <http://www.vmware.com/products/vdi>

Introduction

Sizing the server and storage infrastructure for a virtual desktop infrastructure (VDI) can be a complex task, and, frankly, there are no easy answers when it comes to the tough questions IT architects face on the subject. As part of the sizing process, an architect needs to consider a myriad of factors to ensure a robust end-user experience.

In this paper, we will present a step-by-step process that, if followed, will allow administrators to begin to understand the hardware requirements necessary to build out their VDI implementation.

The key steps in the process are as follows:

- Baseline existing desktop environment
- Estimate VDI hardware needed
- Build proof of concept Infrastructure
- Validate hardware estimates

By the end of this process, an organization should have a very good understanding of how their physical PC environment will map into a virtualized desktop infrastructure and how they should size their VDI implementation to accommodate their targeted users.

As this paper focuses on a methodology that can be applied to a broad spectrum of organizations, it is not a goal to offer a “one size fits all” recommendation on server and storage sizing; given the unique and often times complex nature of any individual customer environment, any “rule of thumb” sizing guidance would be of minimal practical use.

While this document does not provide specific sizing guidance, it will walk through a lab-based case study. This case study will discuss how VMware evaluated a target user group, created the sizing estimates, and validated those estimates in a lab environment. The lab-based case study is focused on Windows XP Professional desktops; however, it is important to note that the process works equally well for desktops running Windows Vista.

After reading this paper, IT administrators should be able to replicate this approach in their environments in order to understand how to scale their VDI hardware assets.

Baseline Existing Desktop Environment

After an administrator has identified the key user groups within the organization that are good candidates for a VDI environment, he or she needs to understand how those users utilize their existing PC resources. The purpose of this step is to understand the performance characteristics of the target users’ workloads – for instance: What applications do the target users need? Are these

applications more CPU- or memory-intensive? Is there an excessive number of storage operations? What type of network load is being generated by the end-user’s activities?

By quantifying the resources being used on a physical desktop, administrators can begin to build a profile of the hardware resources they will need to provide an acceptable end-user experience in a virtualized environment.

Performance Monitoring Tools

The first step in this process is to identify a tool to gather the resource utilization data. Fortunately for administrators, Windows XP ships with the Performance Logs and Alerts tool, otherwise commonly referred to as *Perfmon*. *Perfmon* allows administrators to capture and graph various performance statistics from both local and remote computers.

While it is technically possible to use VMware Capacity Planner to baseline the user environment, historically its applications have been focused in the server arena and it’s not recommended in its current iteration for use with virtual desktops. *Perfmon* was chosen for the case study as it is optimized for desktops and should be readily available on every Windows XP machine.

Key Attributes to Monitor

When looking at resource utilization on the physical desktop, administrators should focus on the following performance indicators:

- **CPU:** Administrators need to understand both average and peak CPU utilization levels
- **Memory:** In the area of memory, administrators should track the amount of RAM being used by the OS and applications
- **Storage:** On the storage front, administrators need to look not only at the amount of storage being utilized, but also at the throughput and input/output operations (IOPS)

The following *Perfmon* counters are suggested as a starting point to help administrators measure the resource utilization on the target desktop:

- CPU Utilization: % Processor Time
- Physical Memory Utilized: Memory Available Mbytes
- Storage Throughput Perfmon: Bytes Transferred / Sec
- Storage I/O Operations (IOPS): Transfers / Sec

Profiling the Target Desktop

After the administrator has chosen the appropriate monitoring tool and counters, the next step is to perform the actual monitoring operation. It is suggested that the administrator run his or her selected monitoring tool over the course of several work days to ensure that normal usage patterns are captured.

For the lab-based case study done at VMware, a typical knowledge worker was selected. This worker’s daily tasks included creating complex documents, presentations, and spreadsheets, accessing the Internet, and e-mail. The user’s PC was monitored over the course of a work week. During this time, the worker utilized fairly standard productivity applications, including Microsoft Office, Internet Explorer, and Adobe Acrobat. At the end of the monitoring period, the data collected were analyzed, and the following values were ascertained:

Resource	Measured Value
----------	----------------

CPU - % Processor Time	2.79%
Memory Consumption – Available Mbytes	245 MB (Total System Memory was 512MB)
Disk – Bytes Transferred / Sec	11760 Bytes
Disk – Transfer Operations / Sec	5 Transfer Operations / Sec

While these values represent the average utilization for each of the listed resources, administrators should also track the peak values as well to help understand the characteristics of their users' workloads. With this information, an administrator can begin the next step in the process – estimating the hardware necessary to host his or her virtualized infrastructure.

Estimate Hardware Needs

The next step in the process is to extrapolate the data collected to get a rough estimate of the number of virtual machines that can be hosted on a single server. By focusing on the capacity of a single server, administrators can define building blocks that can be put in place as their infrastructure scales out.

Before the actual estimates are computed, the IT staff needs to decide which of two approaches to the process they want to use. A server-centric approach entails deciding on an optimal virtual machine density and then computing whether a given server can accommodate the projected workload at the density desired. The workload-centric approach involves looking at the workload that was measured and computing how many discrete workloads can be hosted on a given server.

This paper will focus on the server-centric approach. This approach entails identifying the target number of virtual machines an organization wants to host on a given server and creating an estimate of resources needed based on this desired virtual machine/server density. As a general rule of thumb, densities of 48-64 virtual machines / Server can be achieved. The decision on server density might be driven by concerns around physical space, costs, or the desire to utilize an existing set of VMware ESX servers in the current environment.

Whatever the reason, the administrator will need to develop a baseline number of virtual machines per server that they wish to support and then estimate whether or not the workload they observed in step one can be accommodated. In order to do this, the administrator must perform a series of calculations to extrapolate the data previously gathered on the physical desktop for each component (CPU, Memory, and Storage).

As an example, if cost is a key criterion for an organization, administrators may be willing to trade off performance for a total lower cost per virtual machine. At a basic level, a server configuration that can handle 64 concurrent users will have a cost per virtual machine that is approximately half that of the same configuration running only 32 concurrent users.

For the purposes of this paper, a goal of 64 virtual machines per server was chosen.

Estimating CPU

On the physical PC that was monitored, perfmon showed 2.79% CPU utilization (on a single core 2.2GHz processor); this equated to approximately 130MHz of CPU. Extrapolating this data to 64 virtual machines, a minimum of 8.3GHz of CPU will be needed on the VMware ESX server (64 virtual machines x 130 MHz per virtual machine = 8.3GHZ).

While 8.3GHz represents the amount of processing cycles that will be needed to handle the OS and applications as observed on the physical desktop (based on average values), it is not an accurate estimate. Moving the workload from the physical desktop to a virtual environment will add a measurable amount of overhead. Administrators must consider the additional processing power needed to accommodate the work associated with virtualizing the environment, handling the display protocol, and ensuring that there is enough headroom to accommodate any utilization spikes. By adding an appropriate guard band to the observed processing averages, organizations can help to assure that their implementation can accommodate fluctuations in processor load. As a guideline, the guard band range can be anywhere from 10% (representing a lesser degree of protection from processor spikes) up to 25% (representing a higher degree of protection).

Following on from the above, adding a 25% guard band to the observed 8.3GHz would result in an estimate of 10.38GHz.

Estimating Memory

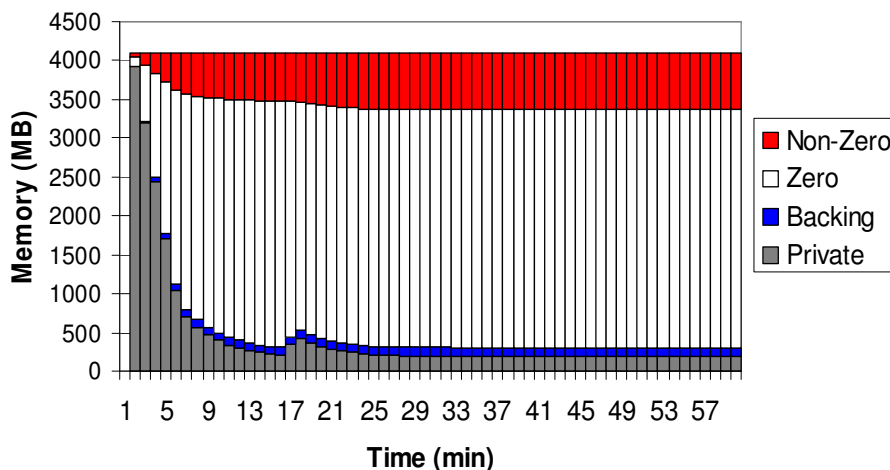
As in estimating CPU, there is not a straight correlation between the amount of physical RAM utilized and the amount of virtual RAM. Unlike with the CPU, where moving into a virtual environment adds overhead to the processing power needed, memory requirements actually decrease when the desktop environment is virtualized. This reduction is a result of a feature in the VMware ESX server commonly referred to as *transparent memory sharing*.

Transparent Memory Sharing

Transparent page sharing in the VMware ESX server is a feature that greatly reduces physical memory requirements. As operating systems and applications are loaded into memory, VMware ESX’s memory manager looks for redundant blocks of code. Rather than have each virtual machine load its own copy of the redundant code, VMware ESX creates a master copy of the memory pages and then redirects each virtual machine to those shared blocks.

This memory sharing can result in each virtual machine requiring less physical RAM than what has been allocated to the virtual machine. Because of this, customers can “overcommit” the RAM in their servers – they are able to assign more virtual memory to their virtual machines than what is housed in the physical servers, allowing for greater scalability and reduced hardware costs.

XP Pro SP2: 4x1GB



The graphic above shows how transparent page sharing affects memory utilization on a VMware ESX server running four 1GB Windows XP virtual machines. The total virtual memory allocated to the four Windows XP virtual machines is 4GB. However, due to transparent page sharing, the physical memory needed by the idle virtual machines, as represented by the gray shaded area, is reduced to only 500MB. This reduction in memory happens over a few minutes as VMware ESX examines memory usage and redirects the virtual machines to use the shared memory pages.

Mapping physical to virtual memory requirements

The targeted user PC utilized, on average, 254MB of RAM. However, in most cases, 512MB is the minimum amount of RAM that should be allocated for a Windows XP virtual machine. Based on 64 virtual machines configured with 512MB of RAM, a total of 32GB of RAM would be needed if transparent memory sharing was not utilized. 32GB would represent the maximum amount of RAM that would be needed to accommodate 64 Windows XP virtual machines.

Based on the graph above, the amount of RAM for Windows XP can be reduced to 125MB per virtual machine. Similar reductions can be observed for common applications. The table below takes average estimates of the optimized memory footprints for a set of Microsoft Office applications and computes the “memory optimized” footprint for a single virtual machine.

Application/Program	Unique Memory Usage
Windows XP	125 MB
Microsoft Word	15 MB
Microsoft Excel	15 MB
Microsoft PowerPoint	10 MB
Microsoft Outlook	10 MB
Total	175MB

Based on the data above, the total memory necessary to support the applications running on Windows XP in 64 virtual machines would be approximately 11.2GB (64 virtual machines x 175MB / virtual machine = 11.2GB).

Adding additional memory to handle spikes and accommodate any extra applications would bring the total estimate to 16GB.

Administrators are faced with a choice of loading anywhere from 16-32GB of RAM in their server. It is recommended to start with the “worst case” scenario (in this example 32GB of RAM) and adjust downward after observing the servers in the proof-of-concept phase.

Storage – Performance

The first facet of storage that needs to be considered in addition to capacity is performance. From a performance perspective, administrators need to look at input/output operations as well as data throughput. The approach to creating the base estimate for storage throughput is pretty straightforward – multiply the observed data from the physical desktop by 64 to compute the values for 64 virtual machines.

IOPS: 5 IOPS x 64 virtual machines = 320 IOPS

Throughput: 115 KBps x 64 virtual machines = 7360 KBps

These estimates need to be evaluated in the context of several other factors. Administrators need to consider whether or not the spindles housing the VDI virtual machines are used for other applications. Additionally, there is a degree of disk I/O overhead associated with the VMware ESX server. Finally, these numbers represent the average values measured on the physical desktop; administrators need to be aware of spikes created by things such as virus scanners and system startup.

Storage – Capacity

Calculating the amount of storage utilized is a fairly straightforward process. Administrators need to take the base size of their proposed virtual machine (accounting for the operating system, applications, and local data) and add enough to account for suspend/resume and page and log files as well as an amount to cover any extra headroom.

In this example, a base .VMDK size of 10GB is assumed. To accommodate the page file and suspend/resume, administrators need to add disk space equal to the RAM allocated to the virtual machine – in this case 512MB. A good estimate for the size of log files is 100MB per virtual machine.

The disk space for a single virtual machine would be as follows:

Base .VMDK:	10GB
Suspend / Resume:	512MB
virtual machine RAM:	512MB
Logs:	<u>100MB</u>
Total:	11.1GB

Multiplying 11.1GB x 64 yields an estimate of 710GB of storage. In order to provide a guard band, it is recommended, based on prior customer experience, to add 15% to this estimate for a total storage footprint of approximately 820GB.

Administrators must also decide on how many virtual machines will be deployed on each LUN. This decision is usually dependent on the particular storage array that will be used.

For VDI environments, some general recommendations are as follows:

- 30-40 .vmdk for average I/O virtual machines
- 15-20 .vmdk for heavy I/O virtual machines

For a more detailed discussion on this topic, please see the following whitepaper:

<http://www.vmware.com/resources/techresources/1059>

It is important to note that these estimates are independent of any storage reduction capabilities. Today's storage vendors provide a myriad of technologies such as deduplication and thin provisioning that can be used to reduce storage requirements. Additionally, the upcoming release of VMware View (the follow-on to VMware VDI) will include VMware View Composer, a software-based feature that can help reduce storage costs by up to 70%.

Estimate Summary

The following table summarizes the estimates for the various configuration parameters based on the data collected from the target desktop user.

Component	Estimate
CPU	10.38 GHz
Memory	17-32 GB
Storage Capacity	2 LUNs @ 410GB Each
Storage IOPS	320 IOPs
Storage Throughput	7360 KBps

It is important to reiterate that many of these estimates do not include the overhead associated with a move to a virtualized environment. The last step in the process will provide an opportunity for administrators to compare their estimates to the actual resource utilization levels.

Build Proof-of-Concept Infrastructure

After an organization estimates the number of virtual machines per core that they wish to support, the next step is to build a proof-of-concept (POC) infrastructure to validate their estimates.

The proof-of-concept environment should be viewed as a microcosm of the desired production infrastructure. To ensure that this happens, a number of questions need to be addressed in the planning phase. This section details some common areas that need to be resolved before building out the proof-of-concept lab.

One of the first considerations that needs to be resolved is around form factor. Architects will need to decide if they are going to use traditional rack-mount or blade servers. Choosing blades will allow higher server densities in the data center but may also increase costs. Administrators also need to consider what requirements they have for expandability and redundancy when they choose their hardware platform.

In the arena of CPU, administrators will need to decide what level of CPU utilization they want to target; it is typically recommended that 65-80% is the ceiling, but some organizations may feel comfortable pushing higher while others will want to leave ample headroom. If the organization wishes to use VMware VMotion (the ability to move virtual machines from one VMware ESX server to another without any downtime for the users of those VMs), they will need to standardize on a single CPU failure. Finally, a cost/benefit analysis should be done on the tradeoffs between dual and quad socket machines; the quad socket machines will allow for a much greater density of virtual machine's per machine; however, the incremental cost associated with these high-end servers may offset these gains.

Any planning exercise needs to ensure that a considerable amount of attention is paid to storage. At the most basic levels, administrators need to decide what protocol they will use for their storage solutions. There are cases to be made for fiber channel, iSCSI, and NFS (as well as direct attached storage) in a VDI implementation; in some cases, companies will want to tier their storage, placing the virtual machines on one class of storage and the user data on another. Architects should also consider whether they will host their infrastructure on dedicated storage boxes or share already existing resources.

Another major decision with regards to storage is the trade-off between drive capacity and rotational speed. When looking at storage components, an architect needs to ensure that the proposed solution meets both the capacity and the storage needs of his or her environment. Any drive class chosen needs to meet both of these needs. An example follows:

The table below lists an example of typical hard drive performance statistics. Always check with your storage vendor for their specific recommendations on disk drive performance statistics.

<u>Drive Speed</u>	<u>IOPS (8KB random mix)</u>
5,400 RPM ATA	50 IOPS
7,200 RPM ATA	60 IOPS
7,200 RPM SATA	80 IOPS
7,200 RPM LCFC	80 IOPS
10,000 RPM FC	130 IOPS
15,000 RPM FC	150 IOPS

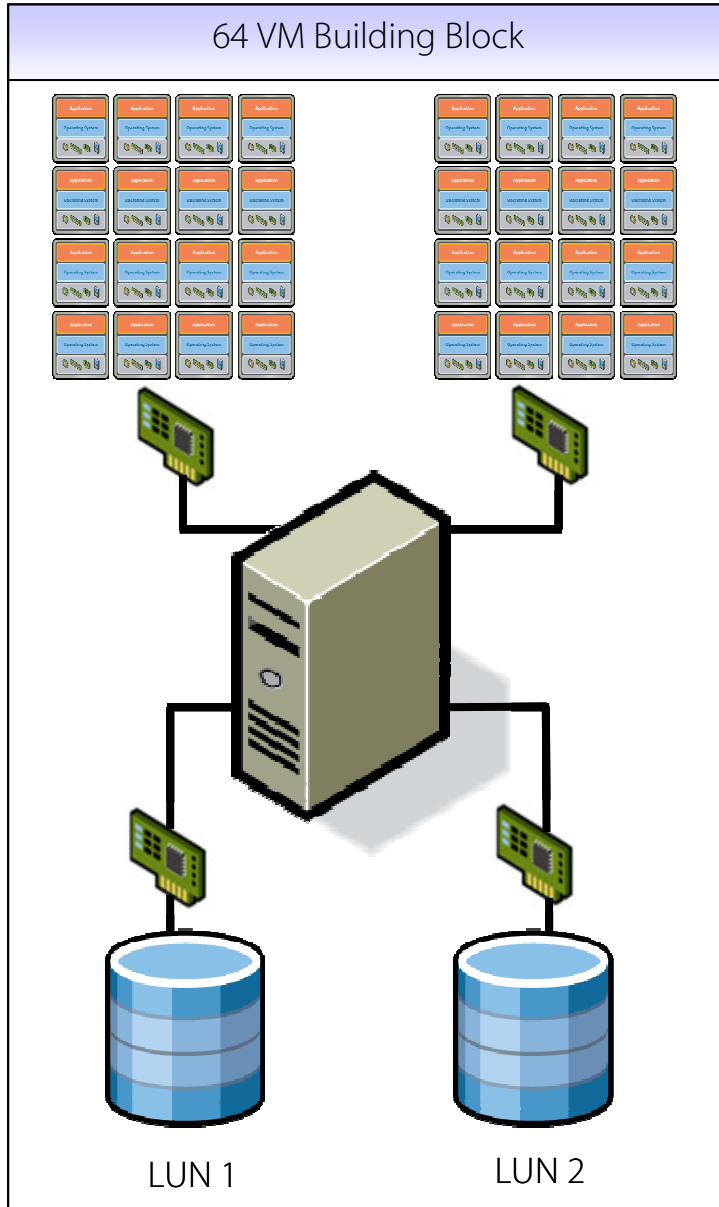
If we return to the observations on storage that were presented earlier in the paper, we will see that we need a set of drives that provides 820GB of storage while delivering 320 IOPS.

Based on the table above, a 146GB 10,000 RPM drive can accommodate 130 IOPS. This means that it would take seven drives (assuming RAID 5) to meet the storage needs ($7 * 146\text{GB} = 876\text{GB}$ usable capacity). However, from a performance standpoint, only three drives would be needed ($3 * 130 \text{ IOPS} = 390 \text{ IOPS}$).

Taking another approach, if 500GB 7,200 RPM SATA drives were used, they could accommodate 60 IOPS. Following the same computations as above, this would indicate that three drives would be needed to meet the storage needs and six drives would be necessary to meet the performance metrics.

All of the considerations discussed previously need to be evaluated in terms of what the company's larger data center and virtualization strategies are. For instance, does the company wish to integrate their VDI solution onto their existing VMware ESX infrastructure, or do they want to implement a dedicated set of hardware for the task? These decisions can be made based on financial, physical, or political factors, and they will almost always be unique to each company.

Lab-based Case Study Test Environment



Based on estimates computed in the previous steps, a proof-of-concept lab was built, and the following hardware was assembled to build out the proof-of-concept for VMware's lab-based case study to host 64 virtual desktops.

The base server config was a dual socket machine running quad-core 2.67 GHz processors equipped with 32GB of RAM. The server ran VMware ESX Server 3.5 Update 2. One server housed the workstation virtual machines while another was used to provide infrastructure services such as Active Directory, and VMware VirtualCenter.

Storage needs were met with an iSCSI array housing 146GB 15,000 RPM SAS drives in a RAID 5 configuration. The array was configured with two LUNs of 410GB each designed to house 32 virtual machines apiece. Tests were performed with both hardware and software iSCSI initiators.

The hosted virtual desktops ran Windows XP SP2 and various applications (including Office 2007). They would be configured with a single virtual CPU and 512MB of virtual RAM.

Finally, all testing was done using VMware VDM 2.1.

Validate Hardware

Estimates

Once the POC hardware has been installed and configured, the final step is to validate the hardware estimates. Organizations can do this either by developing automated test scripts that simulate the workload the targeted users groups would place on the server hardware or by implementing a live evaluation with a targeted group of actual users. It is completely reasonable for a company to utilize both approaches, whereby they perform an initial assessment with automated scripts and then move into a final test phase with live users.

For the lab-based case study performed by VMware, an automated script was developed to simulate the workload of our targeted users. It is important to note that the script was not designed to mimic the actual day-to-day activities of the user, but rather to repeat typical tasks (opening and closing documents, browsing the web, etc) in a manner that created an appropriate amount of stress on the system.

In creating any script, it is important to understand that the workload must be measurable and repeatable. Organizations may want to run several different test passes, and having a predictable work load makes this much easier. Also, any script should try to resemble human activities, but human behavior is very difficult to replicate completely accurately, so the real focus should be on ensuring that the script consumes the same level of resources that were utilized on the physical desktop.

The script used by VMware included the following set of common office activities:

- Microsoft Word – open, modify random pages, save and close the document
- Microsoft Excel – write to random cells, sort data, execute formulas, and create charts
- Microsoft PowerPoint – display slides in slideshow mode, edit slides
- Adobe Acrobat – open PDF document and browse random pages
- Microsoft Internet Explorer 6 -- browse static web pages and web photo album
- WinZip – install and uninstall the application
- McAfee VirusScan – continuous “on-access” scanning

In order to ensure that the script could be tuned to place the appropriate stress on the virtual machine, it was built with a number of configurable parameters. These settings allowed the administrator to adjust the length of time between operations, the number of iterations and the total duration of the execution of the script, and to vary the words-per-minute typing rate. All actions in the script were done on random intervals to ensure that the machine did not fall into a steady state of operation. Finally, in order to monitor the effect of additional virtual machines on end-user performance, in-guest timings were taken. These timings show the rate at which system responsiveness degrades as the intensity of the workload is varied (see below for a more in-depth discussion of this topic).

With the test environment built out and the script written, VMware initiated a series of test runs to validate the sizing estimates that had been made. The following sections present the results of those test runs and compare them to the original estimates.

CPU Observations

The original estimate was that approximately 10.4GHz of CPU would be needed to support the base processing requirements for 64 virtual machines. It is important to remember that this estimate did not include any of the overhead associated with the virtualization of those desktops.

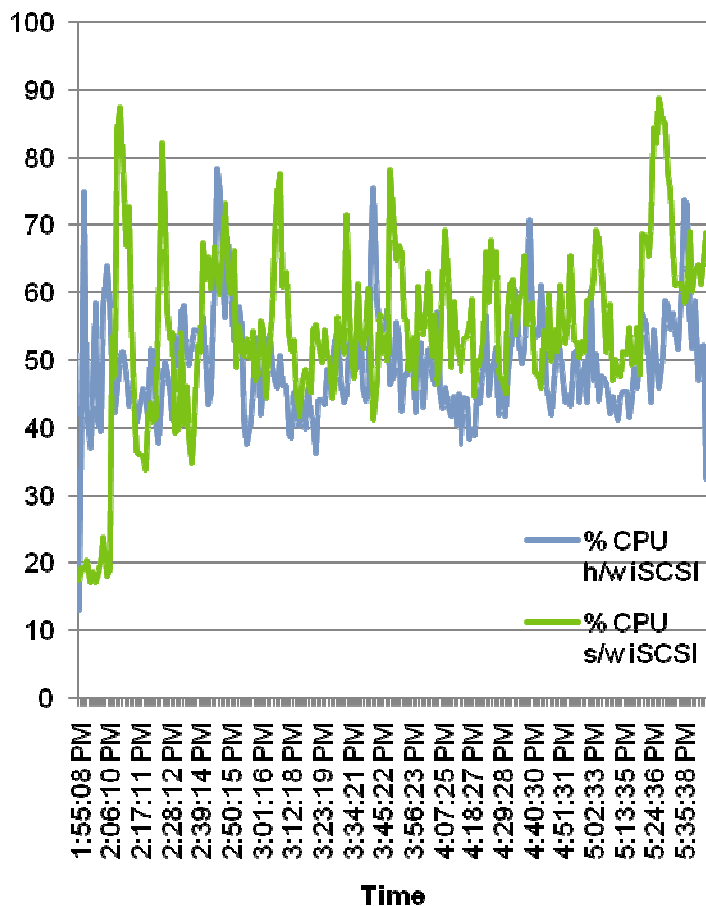
The machines utilized in the test lab had dual quad-core Intel Xeon processors running at 2.67GHz. This offered a total of 21.36 GHz of processing power.

In the testing cycle, VMware ESX was configured to use both hardware iSCSI initiators (represented by the blue line in the graph below) and software iSCSI initiators (represented by the green line in the graph below). The graph shows that average CPU utilization was approximately 6% higher when software-based initiators were used. However, regardless of the initiators, there was still plenty of CPU headroom left, based on the average CPU utilizations of 49% and 55% (for hardware and software initiators respectively).

Based on the available cycles, it appears that 64 virtual desktops utilizing software-based iSCSI consumed approximately 11.75 GHz of CPU, while a hardware-based approach consumed 10.47 GHz of CPU. These numbers indicate that the move to a virtual platform did not significantly increase the CPU power required. It is important to note, however, that these tests did not include the overhead associated with the use of a display protocol. However, given the high amount of CPU headroom available, the addition of a display protocol should not have a major impact on CPU performance.

Finally, it is important to realize that, while the average CPU utilization was well within the estimates, the CPU load itself was fairly bursty – especially in the case of the software iSCSI initiator. Administrators will need to decide if they wish to size for the average utilization or the peaks (see below for more discussion on this topic).

ESX CPU Utilization – 64 VMs



Memory Observations

As previously mentioned, while moving from a physical to a virtual environment results in increased CPU overhead, the opposite is true for memory. And, furthermore, due to memory page sharing, it is difficult to estimate exactly how much memory will be consumed by the virtual machines; our original estimate was that between 17GB and 32GB of RAM would be needed to support 64 hosted virtual desktops.

The graph below shows the memory in megabytes free (represented by the blue line) and the memory in megabytes consumed (represented by the green line) from four different iterations of the test workload.

The graphs show that memory usage grows quickly initially and then flattens out. This is consistent with what would be expected as the page sharing feature kicks in. Page sharing was maximized in this case because all the virtual machines were running a common set of

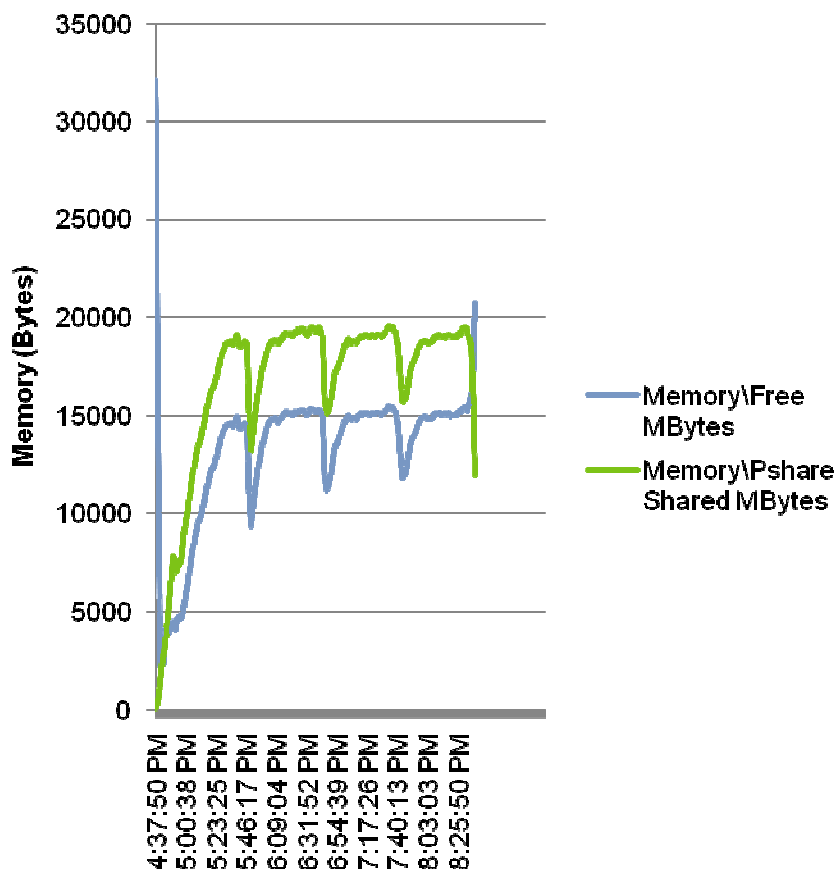
applications. Had each virtual machine been running different apps, then the effects of page sharing would have been diminished.

By examining the graphs, we see that approximately 17MB of RAM is being consumed by the 64 virtual machines, leaving approximately 15MB of RAM free. In reality, administrators could have installed 24GB of RAM (vs. the 32GB that was in the test servers) and still had plenty of headroom for memory spikes.

The concept of dedicating more virtual RAM than what is physically installed in the system is called “overcommitting.” If an administrator had installed 24GB of physical RAM to support 32GB of virtual RAM (64 virtual machines each with 512MB of RAM), we would say they had overcommitted the RAM at a ratio of 1.33:1 (1.33MB of virtual RAM for every 1MB of physical RAM).

By overcommitting RAM, administrators can reduce hardware costs without affecting performance. Ideally, RAM would be overcommitted to the point just before swapping was observed on the system. The key is to ensure that the VMware ESX server is not swapping its memory. This can be monitored through VMware VirtualCenter or esxtop.

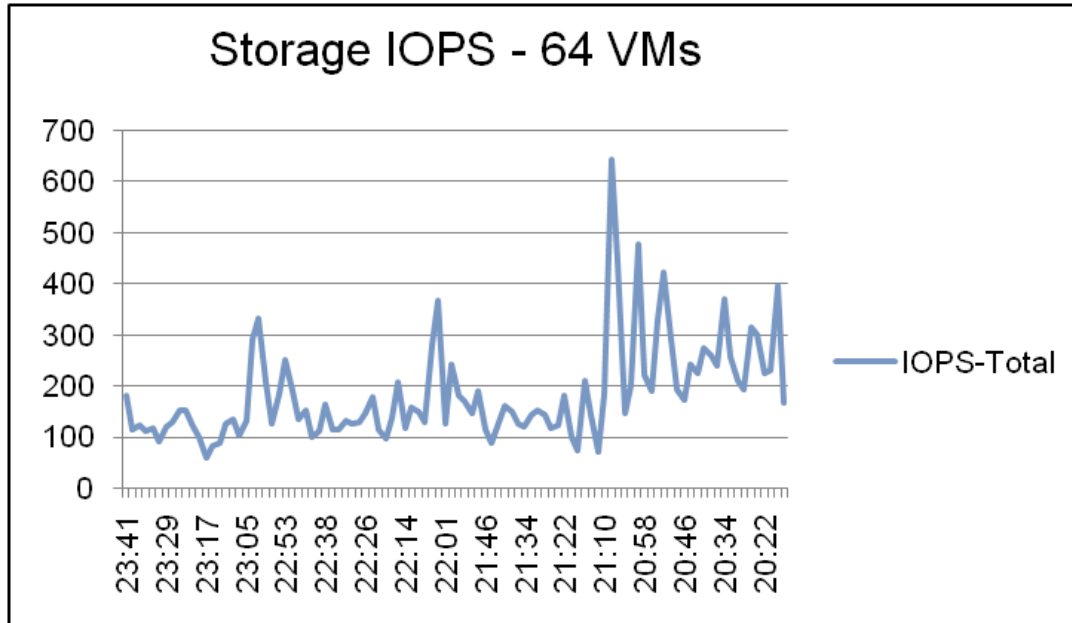
ESX Memory Utilization – 64 VMs



Storage Observations

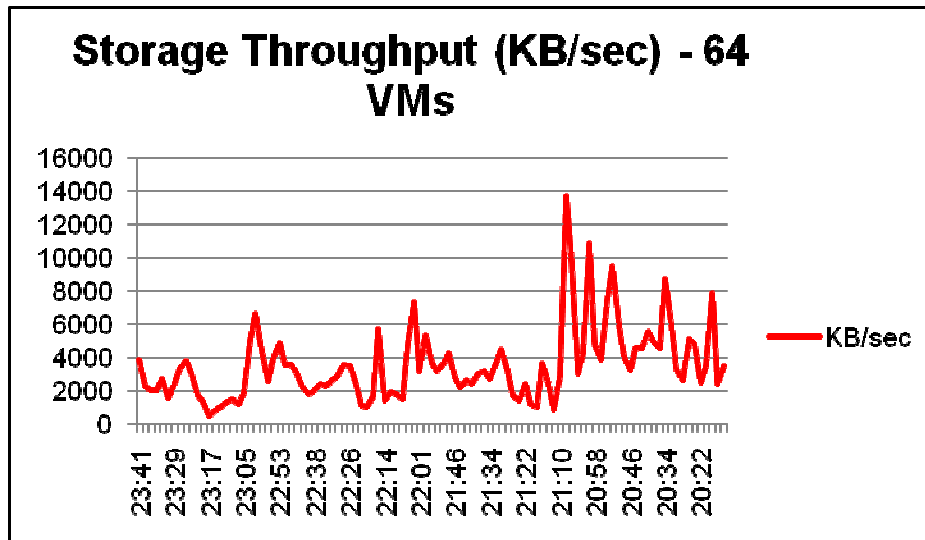
Storage IOPS

The initial estimate for I/O operations on 64 virtual machines was 320 IOPS (input/output operations per second). The graph below shows that the average IOPS was closer to 185, but that the peaks were as high as 650 IOPS. Like with the CPU, the workload showed a high degree of “burstiness.”



Storage Throughput

The characteristics of the throughput readings mirrored the input/output measurements. The average reading from the virtual machines was less than the estimate based on the physical desktops (3530 MB/Sec vs. 7360 MB/Sec), but the peaks were very high – in this case, the highest reading came in at nearly 4x the average (13733 MB/Sec).



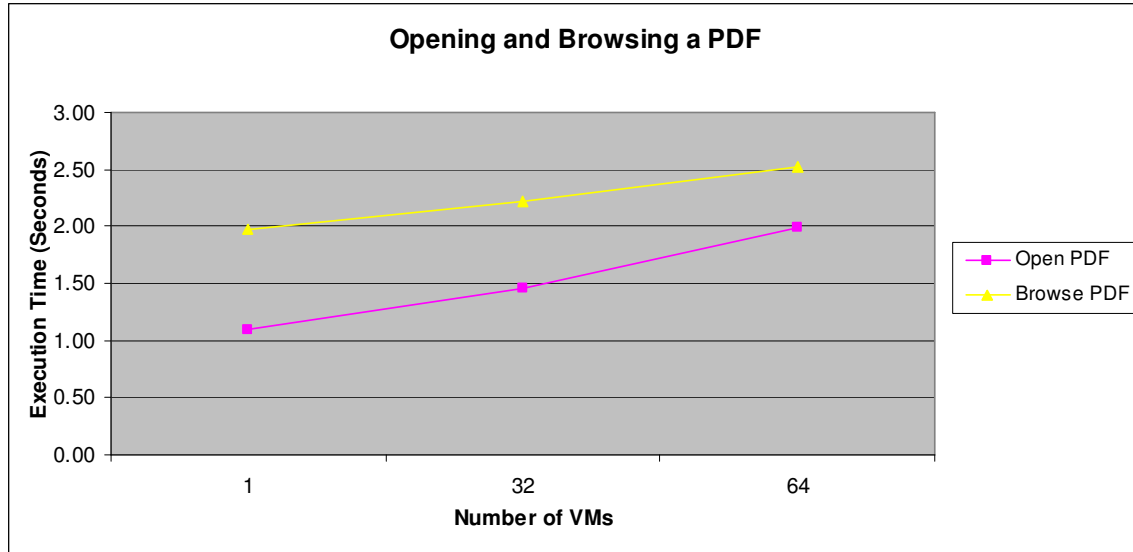
Like with CPU, storage is an area where administrators need to make a conscious decision as to whether they optimize for the average or the peaks, especially when there is such a disparity between the two measurements. Optimizing for the peaks will generally be more expensive, but performance will not suffer if, for instance, a boot or log-in storm occurs. On the other hand, it's significantly cheaper to optimize for the average readings, but administrators could be faced with increased support calls due to poor performance in the event of a log-in or boot storm.

Impact on End-User Experience

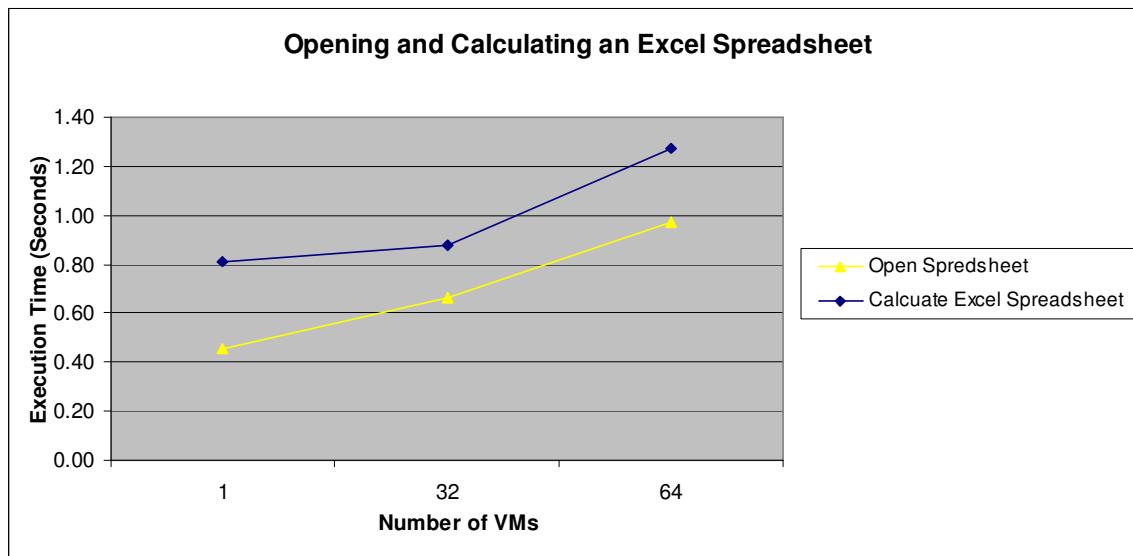
No discussion about scaling would be complete without an examination of the impact of increased workloads on the end-user experience. In reality, administrators could load servers with dozens of virtual machines per core; however, at some point, the system would be so overwhelmed that it would be unusable.

In order to monitor the effect of increased workload densities on end-user performance, timers were inserted into the evaluation script. These timers measured how long various operations – opening a document, saving a document, etc. - took in each virtual machine. A baseline measurement was established by recording the timings of each action for a single virtual machine running on a server. Subsequently, the number of virtual machines per server was increased and the response times tracked. The recorded data showed no significant increase in the time it took the operations to complete as the number of virtual machines per server was scaled from 1 to 64.

The following graphs show how system response for various activities increases as the number of virtual machines running on the server increases.



In the case of both opening and browsing a PDF file, the increase in response time increases in a linear fashion as the number of concurrent virtual machines on the server increases.



Unlike the case with PDF files, where both operations were affected similarly by an increased number of virtual machines, with Excel spreadsheets, the two operations displayed slightly different behavior. As the number of virtual machines increased, there was also a linear increase in the time it took to open the spreadsheet. However, when looking at the time it took to calculate the spreadsheet, there was an increase in execution time of only approximately 10% as the number of virtual machines was scaled from 1 to 32; however, the increase in execution time was closer to 50% as the number of virtual machines was scaled from 32 to 64.

Organizations will need to decide if the increase in execution times is outweighed by the reduced costs associated with having more virtual machines on a given physical server.

In cases where organizations are using live pilots to validate their sizing assumptions, administrators will need to rely on end-user reports to assess the acceptability of the customer experience.

Conclusion

This paper has outlined a set of steps that will allow administrators to determine the appropriate hardware resources to support their VDI implementations. First, administrators need to understand where in their environment virtual desktops make the most sense and then measure the resources those targeted users use today on their desktops. After compiling information on the resource utilization of their physical desktops, administrators can then compute an estimate of the computing infrastructure that will be needed to host the targeted workloads in a virtual environment. Finally, a test bed based on those estimates can be built out to validate the observations.

None of this work can be done in a vacuum. Design decisions need to be made in the context of the trade-offs a company is willing to make. Architects need to decide what the “sweet spot” is with regards to system responsiveness, disaster recovery, end-user experience, and costs.

Unfortunately, the work involved in properly sizing a virtual desktop environment is certainly non-trivial. There is no “one size fits all” answer to the question “how many servers will I need to host my user population.” However, by following a prescribed methodology, architects can begin to design a scalable computing environment that effectively not only meets the needs of their user base but also complies with the IT group’s strategic directions.

Other Resources

VMware VDI Datasheet: http://www.vmware.com/files/pdf/vdi_datasheet.pdf

VMware VDI 60-Day Trial Version: <http://www.vmware.com/download/vdi/eval.html>

VMware VDI Documentation: http://www.vmware.com/support/pubs/vdi_pubs.html

VMware VDI Windows XP Deployment Guide: <http://www.vmware.com/files/pdf/vdi-xp-guide.pdf>

Server and Storage Sizing For VMware VDI: A Prescriptive Approach



VMware, Inc. 3401 Hillview Ave Palo Alto CA 94304 USA Tel 877-486-9273 Fax 650-427-5001 www.vmware.com
Copyright © 2008 VMware, Inc. All rights reserved. Protected by one or more of U.S. Patent Nos.
6,961,806, 6,961,941, 6,880,022, 6,397,242, 6,496,847, 6,704,925, 6,496,847, 6,711,672, 6,725,289,
6,735,601, 6,785,886, 6,789,156, 6,795,966, 6,944,699, 7,069,413, 7,082,598, 7,089,377, 7,111,086,
7,111,145, 7,117,481, 7,149,843, 7,155,558, 7,222,221, 7,260,815, 7,260,820, 7,268,683, 7,275,136,
7,277,998, 7,277,999, 7,278,030, 7,281,102, 7,290,253; patents pending.

VMware Part Number: SG-078-PRD-01-01

