

VMware VDI Storage Considerations



Contents

- Introduction.....2**
- Windows XP Disk I/O Workloads.....2**
- Protocol Choice.....3**
 - Throughput 3
 - VMDK Behavior by Protocol 4
 - Use of Existing Infrastructure..... 5
 - Final Protocol Choice 5
- Storage Array Decisions5**
- Advanced Topics.....6**
 - Thin Provisioning 6
 - Data Deduplication 6
 - Clones 6
- Conclusions7**
- References8**

Introduction

VMware® Virtual Desktop Infrastructure (VDI) is an alternative strategy for providing Windows XP or Vista desktops to information workers. The traditional desktop model involves providing each user with a full set of desktop hardware to run Windows XP or Vista locally: Each user has a full private copy of Windows.

An alternative approach that has been used for the last decade or so is to provide the user with a thin client (either a true thin client or low-end desktop running client software) to access a common server running Microsoft Terminal Services, sometimes with Citrix Presentation Server. In this scenario, each user shares a large Windows Server 2003 instance, with all applications installed on it.

VMware® VDI is essentially a hybrid approach, where each end user has a thin client and connects to a private Windows XP or Vista image—a virtual machine hosted on VMware Virtual Infrastructure. This approach allows IT administrators the greater control over the user environment usually provided by Terminal Services or Citrix environments by consolidating the Windows images on server class hardware. It also allows the images to be stored and managed in the datacenter, while giving each user a full personal copy, which requires no introduction or explanation to a normal user.

This information guide focuses on design choices for the storage environment that forms the foundation of a production VMware VDI implementation.

Windows XP Disk I/O Workloads

For appropriate storage sizing/design decisions to be made, the disk input and output (I/O) characteristics of Windows XP and Vista must be understood. For characterization purposes, workers can be categorized as light or heavy users. Light users typically use e-mail (Outlook), Excel, Word, and web browsing (Internet Explorer or Firefox) during the normal work day. These workers are usually data entry or clerical staff. Heavy users are full knowledge workers using all the tools of the light worker (Outlook, Excel, Word, IE, and Firefox) in addition to working with large PowerPoint presentations and performing other large file manipulations. These workers are usually business management, executives, marketing, etc.

The following table, adapted from the VMware *VDI Server Sizing and Scaling* white paper, compares the disk usage of light and heavy users for a large number of VMware VDI virtual machines (approximately 20) on a single VMware® ESX host. It suggests that over 90% of the average information worker's disk I/O consists of read operations.

	Peak Read Disk I/O	Peak Write Disk I/O	Peak Total Disk I/O
Light User	4.5 MBytes/sec	0.5 MBytes/sec	5.0 MBytes/sec
Heavy User	6.5 MBytes/sec	0.5 MBytes/sec	7.0 MBytes/sec

Table 1. Information Worker Disk I/O Throughput

Before intelligent storage subsystem choices can be made, these throughput values need to be converted to Input/Output operations Per Second (IOPS) values used by the SAN/NAS storage industry. A throughput rate can be converted to IOPS by the following formula:

$$\frac{\text{Throughput (MBytes / sec)} \times 1024 (\text{kbytes / MByte})}{\text{Blocksize (kbytes / IO)}} = \text{IOPS}$$

Even though the standard NTFS file system allocation size is 4k, Windows XP uses a 64-Kbyte block size, and Windows Vista uses a 1-MByte block size, for disk I/O.

Using the worst case (heavy user) scenario of 7.0 MBytes/sec throughput and the smaller block size of 64kbytes, of a full Windows XP group of machines, the generated IOPS for approximately 20 virtual machines is 112 IOPS.

Protocol Choice

VMware ESX 3.0 and later supports multiple protocol options for virtual machine disk (VMDK) storage:

- Fiber Channel Protocol (FCP)
- iSCSI
- NFS
- 10 Gigabit Ethernet (10GbE)

The primary considerations for protocol choice are maximum throughput, VMDK behavior on each protocol, and the cost of reusing existing versus acquiring new storage infrastructure.

Throughput

The following table shows the maximum throughputs of these protocols:

	Maximum Transmission Rate	Theoretical Maximum Throughput	Practical Maximum Throughput	# of VMs on a single host (@7 MBytes/sec for each 20 VMs)
FCP	4 Gbit/sec	512 MBytes/sec	410 MBytes/sec	1171
iSCSI (HW or SW)	1 Gbit/sec	128 MBytes/sec	102 MBytes/sec	291
NFS	1 Gbit/sec	128 MBytes/sec	102 MBytes/sec	291
10 GbE (Fiber)	10 Gbit/sec	1280 MBytes/sec	993 MBytes/sec	2837
10 GbE (Copper)	10 Gbit/sec	1280 MBytes/sec	820 Mbytes/sec	2342

Table 2. Virtual Machines per Host Based on Storage Protocol

The maximum practical throughputs are well above the needs of a maximized VMware ESX server — 128 GBytes of RAM supported in a VMware ESX 3.5 host running 512 MB of RAM per Windows guest.

Hardly any production VMware VDI deployments would use the maximum 128 GBytes of RAM per ESX host supported due to cost constraints, such as the cost of filling a host with 8-GByte memory DIMMs as opposed to 2-GByte or 4-GByte DIMMS. In all likelihood, the VMware ESX host would run out of RAM or CPU time before being bottlenecked on the disk I/O throughput; however, if disk I/O were to become the bottleneck, it would most likely be due to disk layout and spindle count (that is, not enough IOPS). The throughput need of the Windows virtual machine is not typically a deciding factor in the storage design.

NOTE: To present the worst case scenario of one data session using only one physical path, link aggregation is not taken into account.

VMDK Behavior by Protocol

Both FCP and iSCSI are block-level protocols: VMware ESX has direct access to the disk blocks and controls assembly of blocks into files. Block-level protocols are formatted as VMware® VMFS by the VMware ESX hosts and use the VMware ESX file locking mechanism, limited to 32 VMware ESX hosts accessing the same LUN. Block-level protocols also use monolithic, or thick disk, VMDK, where each VMDK is fully provisioned upon creation, so that a 20-GByte disk consumes 20 GBytes of space of the block-level storage regardless of the contents of the VMDK.

NFS is a file-level protocol. The NFS appliance controls the file locking and assembly of blocks into files. File-level protocols use thin disk VMDK format, where a VMDK is only as large as its contents, so that a 20-GByte disk containing 10 GBytes of data consumes 10 GBytes on the NFS storage device. ESX can support up to 32 NFS datastores on a single host.

FCP-attached LUNs formatted as VMware VMFS have been in use since VMware ESX version 2.0. Block-level protocols also allow for the use of Raw Disk Mappings (RDM) for virtual machines; however, RDMs are not normally used for Windows XP or Vista client machines because end-users do not typically have storage requirements that would necessitate an RDM. FCP has been in production use in Windows-based datacenters for much longer than iSCSI or NFS. iSCSI and NFS support were introduced by VMware in VMware ESX 3.0. iSCSI gives the same behavior as FCP, being a block-level protocol, but typically over a less expensive medium (1 GBit/sec Ethernet).

iSCSI solutions can use either the built-in iSCSI software initiator or a hardware iSCSI HBA. The use of software initiators increases the CPU load on the VMware ESX, while HBAs offload this processing time to a dedicated card (as FC HBAs do). To increase the throughput of the TCP/IP transmissions, iSCSI should use jumbo frames. A frame size of 9000 bytes is recommended.

NFS solutions are always software driven; therefore, the storage traffic will increase the CPU load on the VMware ESX servers. For both iSCSI and NFS, TCP/IP offload features of modern network cards can decrease the CPU load of these protocols.

The use of either iSCSI or NFS could necessitate the building of an independent physical Ethernet fabric to separate storage traffic from normal production network traffic, depending on the capacity and architecture of the current datacenter network. FCP always necessitates the use of an independent fiber fabric, which may or may not already exist in a particular datacenter.

Use of Existing Infrastructure

Whether to use existing or acquire new storage infrastructure (fabric and array) should be evaluated based on the capacity and capabilities of any existing devices in the datacenter and determined by the answers to the following questions:

- Is there an existing storage array on the VMware ESX 3.5 hardware compatibility list?
- Does the existing array have enough IOPS capacity for the anticipated number of virtual machines?
- Does the existing array have enough storage capacity for the virtual machines?
- Is there an existing fabric (Ethernet or Fiber Channel) that can support the anticipated number of VMware ESX hosts?
- If there is already a VMware Infrastructure environment for virtualized servers, and, if so, does the storage have enough capacity to support the new VMware VDI environment?

Final Protocol Choice

The final choice of protocol for VMware ESX storage supporting a VMware VDI implementation is often more financial and psychological than technical. If a new fabric and array must be purchased, then the total cost of ownership and return on investment become leading factors in the storage fabric and array decision. If an existing fabric and array can be used, then the new VMware VDI implementation will inherit the technical features of that infrastructure.

Storage Array Decisions

Storage array decisions for a VMware VDI implementation consist of choices for disk type and RAID type. Additional factors often considered include snapshots, clones, replication, and manageability.

Current disk choices are:

- Fibre Channel (FC)
- Serial ATA (SATA)
- Serial Attached SCSI (SAS)

These choices represent the protocol used by the storage array, not between the array and the VMware ESX servers. Disk speed, usually between 7,200 RPM and 15,000 RPM, must also be selected. Disk protocol and speed choices can often be determined by budget, as long as the capacity (in both IOPS and GBytes of space needed) can support the anticipated size of the VDI deployment.

Current RAID level choices are:

- RAID4
- RAID5
- RAID1+0
- RAID6
- RAID-DP (NetApp only)
- MetaRAID's (EMC Clariion only)
- vRAID's (HP EVA only)

The choice of RAID depends largely on the storage array purchased, usually for reasons other than the supported RAID types in a particular mid-tier storage solution. All mid-tier storage arrays can achieve high performance with high redundancy in various ways. As with the individual disk choice,

as long as the RAID choice provides the desired IOPS speed, GBytes capacity, and disk failure redundancy, any RAID listed above can be used.

NOTE: Only 32 VMware ESX hosts can access the same set of 256 block-level assigned LUNs or the same set of 32 NFS datastores.

Advanced Topics

Combining thin provisioning and data deduplication can provide maximum disk usage savings.

Thin Provisioning

Thin provisioning is a storage array function where a LUN or NFS export is presented to hosts at the full size desired, but the array does not reserve the space. Instead, the LUN or NFS export size on disk represents only the contents actually written so far.

Thin provisioning provides value for FCP, iSCSI, or NFS datastores and RDMs. When using thin provisioning, all types of VM ware storage consume disk space only as needed, thereby decreasing the cost of storage for the VMware VDI environment.

Using thin provisioning technology without managing the overall space available on the array presents an inherent danger: since hosts can be told there is more usable disk space on the array than is actually available, it is possible for an array to become full and create host write errors. Use of the monitoring and management software available for a particular vendor's array is highly recommended to avoid this situation.

Data Deduplication

Data deduplication is an algorithm that reduces disk usage. In its most basic form, if multiple storage blocks that contain identical data, then only one copy of the block is stored, and the other locations become place holders, or pointers, to this block. If one of the place holder blocks is changed, the array makes a copy of the source block and applies the change, forming a new, unique block. This operation is seamless and transparent to the VMware ESX hosts consuming storage.

Since the majority of Windows XP or Vista images provided to end users will have very similar, if not identical, disk contents, data deduplication can significantly reduce storage needs. Data deduplication can be seen as a function of the storage array chosen and used with all supported storage protocols (FCP, iSCSI, NFS, and 10 GbE).

NOTE: The data deduplication solution on a given array must support the deduplication of primary in-use storage. Some data deduplication designs reclaim space only during backup operations. This is of little use to a production VDI implementation.

Clones

Multiple vendors and storage arrays have the ability to take a snapshot of a LUN at a point in time and provide it to VMware ESX as a new writable datastore. All changes to the snapshot are written into a different location on the array; however, all reads can access both the changed information and the original datastore. This feature allows the creation of virtual machines a la Linked Clones in VMware Workstation and VMware Lab Manager, where a base read-only disk is used to generate each VM, and all changes are written to delta disks.

Storage savings can be achieved by creating a datastore with, for example, 10 Windows XP images, then copying the datastore as a snapshot and presenting it to VMware ESX 10 times. This presents the VMware VDI environment with 100 virtual machines ready for sysprep, renaming, and deployment. The on-disk usage is that of the original ten VMDK instances, and only the changes are written to the individual virtual machines.

Conclusions

1. Design choices for a production VMware VDI implementation should be based on an understanding of the disk needs of the virtual machines.
Windows XP or Vista clients have radically different needs from virtual machines that provide server functions: The disk I/O for clients is more than 90% read and is rather low (7 MB/bytes/sec or 112 IOPS per 20 virtual machines). In addition, very little disk space is needed beyond the operating system and application installations because all end-user data should be stored in existing network-based centralized storage, on file servers or on NAS devices.
2. Once the disk size, throughput, and IOPS needs of a given VMware VDI deployment are understood, the choices of storage protocol, array type, disk types, and RAID types follow directly.
Thin provisioning, data deduplication, and cloning can drastically lower the on-disk needs in terms of disk space required.
3. The most important consideration in storage decisions may not necessarily be technical but often financial.
Can existing datacenter resources be reused? What is the value proposition and return on investment of acquiring an entirely new storage environment?

References

- Irfan Ahmad, *Easy and Efficient Disk I/O Workload Characterization in VMware ESX Server*, Proceedings of the 2007 IEEE International Symposium on Workload Characterization, Sept. 2007, p. 155.
- A. X. Widmer and P. A. Franaszek, *A DC-Balanced, Partitioned-Block 8B/10B Transmission Code*, IBM Journal of Research and Development, Vol. 27, No. 5, Sept. 1983.
- *Configuration Maximums: VMware Infrastructure 3*:
http://www.vmware.com/pdf/vi3_35/esx_3/r35/vi3_35_25_config_max.pdf
- *Storage / SAN Compatibility Guide for ESX Server 3.5 and ESX Server 3i*:
http://www.vmware.com/pdf/vi35_san_guide.pdf
- *VDI Server Sizing and Scaling*:
http://www.vmware.com/pdf/vdi_sizing_vi3.pdf

Revision: IN-068-INF-01-01



VMware, Inc. 3401 Hillview Ave. Palo Alto CA 94304 USA Tel 650-475-5000 Fax 650-475-5001 www.vmware.com
© 2007 VMware, Inc. All rights reserved. Protected by one or more of U.S. Patent Nos. 6,397,242, 6,496,847, 6,704,925, 6,711,672, 6,725,289, 6,735,601, 6,785,886, 6,789,156, 6,795,966, 6,880,022, 6,961,941, 6,961,806, 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, 7,269,683, 7,275,136, 7,277,998, 7,277,999, 7,278,030, and 7,281,102; patents pending. VMware, the VMware "boxes" logo and design, Virtual SMP and VMotion are registered trademarks or trademarks of VMware, Inc. in the United States and/or other jurisdictions. All other marks and names mentioned herein may be trademarks of their respective companies.

