VMFS File Locking and Its Impact in VMware® View™ 5.1
Table of Contents

Introduction ................................................................. 3
  About VMware View ...................................................3
  About This Paper ......................................................3
Large-Scale Pool Support for View 5.1 .................................... 4
  Background ..............................................................4
    Full Clones and Linked Clones ................................. 5
    Linked Clone Pool Provisioning in VMware View .......... 6
    Parent Image and Replica ........................................6
  Scalability Issues in VMware View Deployment ......................7
VMFS Lock Mechanism .................................................. 8
  Exclusive Lock ........................................................9
  Read-Only Lock ......................................................9
  Multi-Writer Lock ..................................................10
Why View Is Affected by This Limit ........................................10
Workarounds for VMFS Limitations .......................................11
  Limit Cluster Size to Eight Hosts .............................11
  Base Image Duplication .........................................11
  Use NFS Storage for Replica ....................................11
Implementation in View 5.1 ............................................12
Use Cases Using a Large Cluster with VMware View 5.1 ...............14
  Use Case 1 – Replica Stored in Separate Datastore ..........14
  Use Case 2 – Replica Stored with OS Disk ......................14
  Use Case 3 – User Data Disk Stored in Separate Datastore ...14
  Use Case 4 – Increasing the Cluster Size to More than Eight Hosts ..........14
  Use Case 5 – Reusing Old Machines ...........................14
Advantages and Disadvantages ...........................................15
  Advantages .........................................................15
  Disadvantages .....................................................15
Summary .................................................................16
  References ..........................................................16
About the Author ........................................................16
  Acknowledgments ................................................16
Introduction

About VMware View

VMware® View™ is a desktop virtualization solution that allows IT organizations to overcome the hassles of traditional server-based computing, without the challenges that often accompany server-based solutions. By leveraging the benefits of VMware View, organizations can take the first steps toward transitioning away from the era of distributed PC computing and moving toward Cloud Computing and the delivery of user desktops as a service. Built on the industry-leading virtualization platform VMware vSphere™, VMware View is a solution that lets you manage operating systems, hardware, applications, and users independently of each other, wherever they may reside. VMware View streamlines desktop and application management, reduces costs, and increases data security through centralization, resulting in greater user flexibility and IT control. View 5.1 is the latest View release by VMware and it improves scalability by introducing several new features.

About This Paper

View versions prior to 5.1 had a limitation of supporting a maximum of eight hosts in an ESX cluster when VMware® vStorage Virtual Machine File System (VMFS) storage was used. This was a barrier in a large-scale deployment of virtual desktops. View 5.1 breaks this scalability limitation by supporting a maximum of 32 hosts in an ESX cluster provided NFS storage is used for linked clone replica creation. Some common questions from customers include:

- What are the technical reasons for the eight-host limit?
- Why is View affected by this limit while other VMware products are not?
- Are there workarounds to overcome this limit?
- How does View 5.1 overcome this limit?

This paper throws some light on these topics and tries to answer similar questions, in addition to presenting a few use cases using View with a cluster having more than eight hosts.
Large-Scale Pool Support for View 5.1

VMware View 5.1 extends provisioning scalability by supporting a maximum of 32 hosts in an ESX cluster, provided an NFS datastore is specified for linked clone replica creation. This feature provides more flexibility for Virtual Desktop Infrastructure (VDI) design considerations and offers greater manageability by reducing the required number of clusters in the environment.

Background

Typical vSphere cluster configuration includes multiple ESX hosts connected to a shared storage (datastore). The aggregate computing resource of the ESX host farm is referred to as a 'cluster.' Virtual machines are stored in the shared datastore and get powered on in one of the ESX hosts in this cluster. Since the same datastore is connected to all the ESX hosts, the host in which the virtual machine needs to be placed and powered on can be selected dynamically by Distributed Resource Scheduling (DRS). DRS ensures that virtual machines are evenly distributed across the hosts so as to balance the resource utilization on each host.

Figure 1: Distributed Resource Scheduling Balances Resource Utilization

This vSphere infrastructure is the primary back end of many other VMware solutions like VMware View, VMware vCloud™ Director, etc. In these solutions, large numbers of virtual machines are deployed in seconds and managed via an administrator interface. When a large number of virtual machines with exactly the same configuration needs to be created, VMware leverages cloning technology. Cloning is the technique where a master virtual machine is created and used as a base image to create any number of virtual machines that have the exact same characteristics of the master virtual machine. There are two types of cloning—full cloning and linked cloning.
Full Clones and Linked Clones

A full clone is an independent copy of a virtual machine. A full clone shares nothing with the parent once the cloning operation is complete. Once a full clone is created, all the operation associated with it is separate from its parent.

A linked clone is a copy of a virtual machine that shares the virtual disks with its parent. The base disk will be shared by many linked clones, and any changes made to the linked clones after creation go into another small disk known as a redo disk or delta disk. Since the base disk is shared by many linked clones, a linked clone must always have access to the base disk; without the base disk it becomes unusable. Each linked clone occupies a smaller virtual disk space than its parent. Linked clones access software installed on the parent thereby saving large amounts of storage space, a major consideration in large deployments.

Each linked clone acts like an independent virtual machine with a unique identity, including a unique hostname and IP address, but requires significantly less storage than a full clone. Practically unlimited numbers of linked clones can be created from one parent image, reducing the total storage space required.

VMware View uses software called VMware View Composer which is capable of creating linked clones from a base parent image snapshot. This may be installed along with the VMware vCenter™ server or it can be in a separate virtual or physical machine.

![Linked Clone Layout](image-url)

*Figure 2: Linked Clone Layout*
**Linked Clone Pool Provisioning in VMware View**

In order to provision a linked clone desktop pool in VMware View, the following steps need to be performed. It is assumed that a working VMware View environment is ready with vCenter Server added and View Composer installed.

- Create a master virtual machine with a desktop OS installed
- Take a powered-off snapshot of the master virtual machine
- Open the View Administrator Console and invoke the automated linked clone pool creation wizard
- Choose default options and select the prepared master image and snapshot
- Select the cluster or host and datastores to complete the pool creation

On completion of the wizard, pool provisioning starts creating one or more full clones of the master virtual machine known as replicas.

**Parent Image and Replica**

Linked clones created from the parent image with VMware View are linked to a full clone virtual disk called a replica. Replicas are created on a per-datastore basis, so that each datastore used for desktop creation and storage carries a replica.

In VMware View, there is an option to store all the replicas together in a separate datastore. If a separate datastore is specified for replicas during pool creation, all the replicas will be created in the replica datastore only, and no replica creation will be performed on a per-datastore basis.

The replica is cloned from the powered-off snapshot of the master virtual machine. If a new snapshot is taken for the master virtual machine and a pool is provisioned using that snapshot, a new replica will be created. When you snapshot a virtual machine, the virtual disk is set to read only (parent disk) and an additional virtual disk is chained to the read-only disk where the virtual machine can issue additional write operations. If you create a second snapshot, the first chained disk also becomes read only and a second disk is created and chained to the first snapshot disk to take the write operations of the virtual machine after the second snapshot was taken. The last disk in this chain is called the leaf node.

For VMware View, the replica is the base disk for all the linked clones associated with it, so it is important that the replica is not corrupted or deleted. The replica virtual machine will be write protected and it cannot be deleted from the View Composer. View provides an option to place all the replica disks on a high-performance storage in order to improve I/O operations between the replica and the associated linked clones. The linked clone mechanism makes it easier to install patches by updating the parent virtual machine, taking a new powered-off snapshot, and recomposing the pool to the new snapshot.
Scalability Issues in VMware View Deployment

As we discussed earlier, VMware View is a solution to deploy and manage large pools of virtual desktops. A typical View environment can contain hundreds or thousands of virtual desktops. Typically these virtual desktops will reside on a shared VMFS datastore that will be accessed by several ESX hosts. In order to host the large numbers of virtual machines that are part of a single pool, customers will need a considerable number of ESX hosts in a single cluster.

The ESX server uses VMFS to host virtual machines. There is a limit for VMFS where only eight ESX hosts can concurrently share a virtual disk.

The eight-host limit is a VMFS limit at the file-system level where each tree of linked clones can only be run on eight ESX servers. For instance, if there are a hundred linked clone disks off the same replica with several leaf nodes in the tree, all leaf nodes can be run simultaneously—but they can only run on up to eight ESX hosts. A different tree can run on a different set of eight hosts. The situation is caused by file locking in VMFS.
VMFS Lock Mechanism

File locking is a mechanism that restricts metadata access to a virtual disk by only one entity at a specific time. When two processes from different hosts access a single file, there can be a harmful collision between processes or threads that share these files. Consider an example where two hosts, Host A and Host B, access a shared virtual disk at the same time. Both these hosts currently have the state of the virtual disk, which we will call State 1. Suppose some configuration changes are made to Host A, and the changes are written into the virtual disk to bring it to State 2. If Host B, which still has the original stale state (State 1), writes back the same value to the virtual disk, the virtual disk returns to State 1, losing the changes made by Host A.

Figure 4: File Locking

File locking prevents this problem by enforcing periodic updates to the file. VMFS uses on-disk lock for synchronizing concurrent access to shared resources like files and directories from multiple ESX hosts that are connected to the shared storage. VMFS makes use of three different types of locks—exclusive lock, read-only lock, and multi-writer lock.
Exclusive Lock
Exclusive locks exclude all except one host from accessing the associated resource of the lock. An exclusive lock allows only one process/host to update a particular piece of data (insert, update, and delete). When one process has an exclusive lock on a row or table, no other lock of any type may be placed on it. In the shared file system master/slave pattern, there is nothing special to distinguish a master host from the slave hosts and hence all the hosts are considered equal. There can be any number of hosts in a VMFS cluster. The hosts in the cluster therefore compete to grab the exclusive lock on the data file. If Host B has acquired the exclusive lock in a multi-host cluster, the hosts behave as follows:

- Host B retains the exclusive lock on the data file, preventing the other hosts from accessing the data. Host B starts up its transport connectors and network connectors, enabling other messaging clients and message brokers to connect to it.
- The other hosts in the cluster keep attempting to grab the lock on the data file, but they do not succeed as long as Host B is running. Other hosts do not start up any transport connectors or network connectors and are thus inaccessible to messaging clients and brokers.

Read-Only Lock
Read-only locks are used to grant read-only access to concurrent hosts accessing the resources associated with the lock. More than one host can access the disk but none of the hosts can modify it. If a host acquires a lock in read-only mode, the lock may also be granted to other hosts that request the lock in read-only mode; the lock can never be granted to hosts that request the lock in exclusive mode or multi-writer mode.
Multi-Writer Lock
Mult-writer locks are used to grant concurrent writing privileges to multiple hosts. In a cluster having multiple ESX hosts, each ESX server stores its virtual machine files in a specific subdirectory on the VMFS. When a virtual machine is operating, VMFS has a lock on those files so that other ESX servers cannot update them. VMFS ensures the virtual machine cannot be opened by more than one ESX server in the cluster. VMFS provides on-disk locking that allows concurrent sharing of virtual machine storage resources across many ESX nodes. Furthermore, VMFS manages storage access for multiple ESX instances and allows them to read and write to the same storage pool at the same time. If a host acquires a lock in multi-writer mode, the lock may also be granted to other hosts that request the lock in multi-writer mode.

Why View Is Affected by This Limit
VMFS, being a clustered file system, requires using read-only and multi-write locks since files are shared by many hosts at the same time. Files opened in read-only and multi-write locking mode cannot be accessed by greater than eight hosts, and the same limitation applies to snapshot hierarchies and linked clones.

In case of read-only locks and multi-writer locks, there may be multiple hosts that hold the lock at the same time. In order to keep track of all these hosts we need to place the host User IDs (UIDs) in the lock structure. The host UID is a unique identifier that identifies the ESX host that owns the lock at a given point in time. There is a field in the lock structure called ‘holders’ where we place the host UIDs. To update the lock automatically, we need the lock to fit in a single sector. Therefore the on-disk lock needs to be 1 sector in size. The size of this structure is 512 bytes. Since the lock structure on disk is only 512 bytes (1 sector) we can only store a limited number of host UIDs. This limit is currently eight. As a consequence, we can only grant the lock to eight hosts in read-only and multi-writer mode. Subsequent access to the resources will fail with an error.
Since VMware View is designed for provisioning of linked clones, it is affected by this limit. Eight hosts can be a lot of capacity for server-oriented workloads, so it’s not an issue in many use cases. But this can be an issue when low-end servers are used. We may need to use a large number of low-end servers to set up a medium-size virtual desktop pool since hardware resources are limited. Additionally, when we need to deploy very large virtual desktop pools, we may need to use more than eight hosts of good configuration.

VMware View may work fine most of the time with more than eight hosts per datastore as long as you aren’t using DRS on your cluster. But this configuration is prone to several errors, and hence not advisable. Suppose we have a cluster with nine hosts in a View environment. We already have linked clones sharing the parent disk powered on in all eight hosts. If DRS places a linked clone with the same replica on the ninth host, the virtual machine will fail to power on. Further, if we have CD and floppy images attached to virtual machines deployed on more than eight hosts, they will be inaccessible when the ninth host comes into the picture.

**Workarounds for VMFS Limitations**

The eight-host limit is affected only by products which make use of linked clone technology to share a base disk by the linked clones. Examples are VMware View and VMware vCloud Director. In View, the limit is applicable only for linked clone virtual machines, and not for full clone virtual machines. There are workarounds to overcome this limit, but usually, these workarounds increase management complexity and limit the utility of the features like DRS. These workarounds include:

**Limit Cluster Size to Eight Hosts**

This is the most commonly used and easily adaptable workaround to eliminate the eight-host limit issue. But this is possible only in the case of small-scale deployments. When large-scale deployments are considered, this workaround impacts scalability and increases management complexity.

**Base Image Duplication**

In this method, the base images are duplicated for each set of eight hosts and affinity rules are set to control DRS placement. This is very expensive in terms of storage consumption and implementation is complex. Base image duplication should be configured using third-party tools and this does not come as a feature in View.

**Use NFS Storage for Replica**

This is a new feature implemented in VMware View 5.1 to support clusters having more than eight hosts. This will be possible only in environments that have NAS storage in place.
Implementation in View 5.1

There are two types of file locking mechanisms—mandatory locking and advisory locking.

Mandatory locking does not require co-operation from the processes accessing the same file. Mandatory locking asks the kernel to check each and every file operations line—open, read, and write—to verify that the process is not violating the lock on the shared file.

Advisory locking requires co-operation from the processes accessing the same file. For example, if process A acquires a write lock and started writing into the shared file, and process B, without trying to access the lock, opens the file and writes into it, process B is a non-co-operating process. With an advisory lock system, processes can still read and write from a file while it’s locked.

NFS protocol by default does not support file locking, but it supports a protocol called Network Lock Manager (NLM). The Network Lock Manager uses an advisory locking scheme for locking.

In this figure, clients A and B use advisory locking to coordinate their changes in the data file. The read and write operations will not proceed unless they successfully establish an advisory lock. The advisory lock does not prevent client C from ignoring the lock and it is capable of reading from or writing to the locked file. This makes NFS capable of sharing a media or CDROM among several machines at a time and this property of NFS is utilized in VMware View 5.1.

The eight-host limit is the property of the VMFS file system and network attached storage (like NFS) is not affected by this limit. But VMware View releases prior to View 5.1 did not allow selection of a cluster having more than eight hosts using linked clone pool creation irrespective of the storage type used. In View 5.1 we are relaxing this restriction by allowing a user to select a cluster having more than eight hosts, provided NFS storage is used for placing the replicas. The linked clones can still be placed in the VMFS storage.
The following figure shows the storage selection UI in View 5.1 where a large cluster is used, and VMFS storage is filtered out. Clicking the filtered out storage brings up a pop up explaining that VMFS storage cannot be used since the cluster contains more than eight hosts.

Figure 10: Storage Selection UI in View 5.1
Use Cases Using a Large Cluster with VMware View 5.1

Following are a few use cases and best practices that should be followed while using a large cluster (having more than eight hosts) in VMware View 5.1.

Use Case 1 – Replica Stored in Separate Datastore
In this case, during pool creation if you select a cluster having more than eight hosts, the VMFS storage will be filtered out for placing a replica and you must use an NFS datastore for a replica. You will still be able to select VMFS for placing an OS disk and user data disk.

Use Case 2 – Replica Stored with OS Disk
In this case, during pool creation if you select a cluster having more than eight hosts, the VMFS storage will filter out OS disks (since the replica is stored along with the OS disk) and you must use an NFS datastore for OS disks. You will still be able to select VMFS for placing a user data disk.

Use Case 3 – User Data Disk Stored in Separate Datastore
In this case, during pool creation if you select a cluster having more than eight hosts, the persistent disks can be stored in an NFS or VMFS datastore.

Use Case 4 – Increasing the Cluster Size to More than Eight Hosts
Suppose there is a View environment using a cluster having less than or equal to eight hosts, and the administrator wants to increase the cluster size by adding more hosts or wants to move to a new cluster having more than eight hosts. If the replica is already stored in NFS storage, no action is required. But if the replica is in VMFS storage, and the cluster size is increased to more than eight hosts, further provisioning and pool maintenance operations will fail. The View administrator console as well as event logs will show an error saying that the cluster has more than eight hosts. In order to avoid this, the following steps need to be performed:

• Edit the pool and navigate to the vCenter Settings page.
• If the replica is stored in a separate datastore, deselect the datastore and select an NFS datastore for the replica.
• If the replica is stored along with the OS disk, deselect the datastore for the OS disk and select an NFS datastore.
• Perform a rebalance operation for the pool.

Use Case 5 – Reusing Old Machines
Addressing the eight-host limit in View will be most useful when an organization plans to reuse their existing servers. There may be old servers with lower-end configurations and if a moderate number of these machines can be put together in a cluster, there is a possibility of provisioning large pools. Since VMFS can support up to 32 hosts, the organization can combine 32 hosts together in a cluster if they are using NFS storage for placing replicas.
Advantages and Disadvantages

In order to support a cluster with more than eight hosts with VMware View, NFS storage should be used. NFS storage has several advantages, and a few disadvantages, some of which are listed below:

Advantages

• The major advantage of NFS storage is that the files can be shared among many hosts allowing simultaneous access. This is the reason for the use of NFS storage when we deal with clusters having large number of hosts.

• Rack-mounted servers used to have a limited number of slots available for adapter cards and this limited fiber channel storage use. Since NFS uses an Ethernet- and IP-based solution, it can effectively use the embedded gigabit Ethernet ports available on the server motherboard.

• It is very easy to add new a NFS datastore since it is as simple as mounting a file system, rather than creating LUNs as in the case of fiber channel or iSCSI datastores.

• If your deployment includes NAS devices that support the vStorage APIs for Array Integration (VAAI), you can enable the View Composer Array Integration feature on linked-clone pools. This feature uses native NFS snapshot technology to clone virtual machines. Native snapshot is a technology where a snapshot is taken at the storage level rather than server level. With this technology, the NFS disk array clones the virtual machine files without having the ESXi host read and write the data. This operation might reduce the time and network load when virtual machines are cloned. This will ensure that server resources are not utilized and server performance is unaffected. Native snapshot technology can be configured only with NAS storage like NFS. Please note that the VAAI feature is supported on vSphere 5.0 and later and you cannot use this feature if you store replicas and OS disks on separate datastores.

Disadvantages

• I/O performance needs to be considered while using NFS storage, because some industry observers counter that NFS does not have the same performance characteristics compared to other types of storage. Under high network traffic, NFS performance might slow down.

• Since NFS is based on Remote Procedure Call (RPC), it is sometimes less secure and should only be used on a trusted network or behind a firewall.
Summary

The objective of this white paper is to provide information on the eight-host limit in VMFS and how VMware View is affected by this limit. This paper gives some insight on the history and technical details of this limit. It elaborates on the reason for such a limit in releases prior to VMware View 5.1, and explains how the limit is removed in View 5.1. It also lists a few use cases for using View 5.1 on a larger cluster and the advantages as well as disadvantages of using NFS storage in a View environment.

References

For more information about VMware View 5.1, please visit the product pages:

• Product Overview
• Product Documentation
• VMware View 5.1 Documentation Center

For more information about Network File systems, please visit

• Network File System (NFS)
• Linux NFS – HOWTO

For more information about VMFS, please visit

• VMFS Documentation
• VMFS Best Practices

About the Author

Jubish Jose is a Member of Technical Staff in the End User Computing group at VMware. Being part of the R&D team, he is an expert in VMware View, especially in features involving a VMware View Composer component.

Follow Jubish at http://communities.vmware.com/people/Jubish

Acknowledgments

VMware would like to thank the following individuals for their contributions to this paper and help with content review:

R&D Engineering - Frank Taylor, Mikhail Chainani, and Mayank Rawat
Product Marketing - Fred Schimscheimer