Technical report:

VMware ESX Server 3.0 on IBM System Storage N series

Best-practices backup case study

Document NS3562-0

April 11, 2008

Table of contents

Abstract

This document discusses how IBM System Storage N series and VMware ESX Server 3.0 work together to provide an effective backup solution. The document showcases a best-practices case study involving the Loyola Marymount University.

Introduction

The goal of a VMware and IBM® System Storage™ N series solution is to provide a crash-consistent process that takes advantage of snapshot-based backups.

The solution is based on the IBM System Storage N series with Snapshot™ technology's hot snapshot methodology. Because the solution originally was designed for ESX 2.5, the reader should be aware of the following important design and syntax considerations between ESX 2.5 and 3.0, which this paper and best-practice case uses:

- Between ESX 2.5 and 3.0, VMware changed the script command syntax.
- In ESX 3.0, it is now necessary to account for Vmotion and the fact that the location of a virtual machine (VM), and therefore the target host for snapshot commands, can change.
- The storage for each VM is a separate raw device mapping (RDM) logical unit number (LUN), and all RDM LUNs are in one N series flexible volume.

The decision to use RDMs instead of VM disk format (VMDK, or .vdmk) was based on the desire for file-level restore capability. With either topology, it is possible to mount the backup copy as a second disk on a helper VM. However, with RDMs, it is also possible to use IBM System Storage N series with SnapDrive® to mount the backup copy on a physical server. Also, VMware recommends RDM for situations where storage area network (SAN) snapshots are desired.

RDM of a VM virtual disk is used for some hardware snapshot functions of the disk array or to access the disk from both a VM and a physical machine in a cold standby host configuration for data LUNs.

Background on technical issues

Although VMware changed some names and commands between ESX 2.5 and 3.0, the concepts and functionality are the same. Taking a VMware snapshot, whether from the graphical user interface (GUI) or the command line, does indeed quiesce the vmdk.

In the case of RDM LUNs, the metadata and change log for the snapshot are stored in the VM file system (VMFS) volume where the VM configuration file (VMX or .vmx) resides. Restores can recover to the point in time when the VMware snapshot was taken. Even without quiescing the .vmdk, the SAN snapshot copy can be crash consistent. For a Microsoft® Windows® VM, Checkdisk can run if you boot from that state.

In ESX 2.5, the change log was called a REDO log. ESX 3.0 calls them copy-on-write (COW) files. COW files store a copy of each changed block in 16MB increments. For multiple VMware snapshots or VMs with heavy disk I/O, the COW files can grow large quickly. However, for the purpose of taking a SAN snapshot, this shouldn't be a concern because the snapshot will exist only for the duration of the backup script execution.

Note: The COW files do not have a ".cow" extension. They appear as the original name of the VMDK file with a sequence number appended, such as "filename_0001_vmdk."

There may be a concern with the number of simultaneous VMware snapshots on a single VMFS volume on the SAN, with multiple ESX servers connecting to it. According to the VMware Consolidated Backup (VCB) design team, it is not good to have too many simultaneous disk writes to VMFS from multiple ESX servers. VMware had suggested that more than 50 VMs on the same VMFS volume (all with VMware snapshots) might cause performance degradation. Data replication service (DRS) affinity rules or other grouping criteria in SQL can be used to group VMs when running the quiescing script.

In addition, when automating VMware snapshots, it should be noted that the host ESX server for a particular VM can change with VMotion, DRS, and high availability (HA). The script documented in the Appendix must send the quiescing command to the correct host, but may not know which host has a particular VM.

N series and VMware impact

Within ESX, there are many possible methods to perform a backup to a storage system. It is important to use a deterministic strategy to perform a backup and recovery analysis in light of the desired backup and recovery points. This should be done according to customer approved service level agreements (SLAs) or goals in order to come up with a viable solution to minimize downtime and address these goals.

This document presents a storage-based backup solution for RDMs, which can be fully accessed via Fibre Channel (FC) or iSCSI. The method presented in this white paper produces the following benefits:

- Zero downtime
- Zero CPU load on host and guest
- Zero network traffic
- Instantaneous
- 24/7 backup window
- No need for backup agent in guest.

Systems administrators who have experienced problems with backing up VMs as physical servers have chosen to back up their VMware environment by backing up the files that make up the VM (the virtual disk files and configuration files). Backing up this data directly to tape drives results in the same problems as those encountered for traditional file-based backups. There is typically too much data behind each physical server to back up in a traditional backup window.

To maintain high utilization ratios, many customers have asked their storage vendors to implement some form of storage-based backup for their virtual infrastructure. With this method, the virtual machines are placed in a hot backup mode; the virtual disks are locked and all new data is written to temporary log files. Once in this state, the virtual disks are backed up. When the virtual disks have been successfully backed up, the locks are released and the contents of the temporary files are flushed back into the virtual disks.

Disk-based backup practices include copying the VMDK file from the production disk to a second set of disks, or, for customers who want a faster operation, some form of split mirror backup technology. Although both of these solutions provide a much faster backup than backing up directly from the production system to tape, both solutions require 100% additional storage for every backup, and that storage needs to be completed and kept online. This requirement for additional storage runs contrary to the utilization goals associated with VMware deployments and should not be considered. Some storage vendors offer copy-out snapshot technologies as alternatives to the 100% additional storage required with split mirror technologies. The I/O overhead required with copy-out snapshot technologies, and the subsequent performance impact, prevent these solutions from being implemented.

The inherent negative features of traditional disk-based backups do not apply to the patented N series Snapshot technology. With N series technology there is no performance penalty for taking snapshot copies, because the data is never moved, as it is with copy-out technologies.

The cost for snapshot copies is only at the rate of block level changes, not 100% for each backup, as with mirror copies. By combining N series Snapshot technology with VMware ESX server, administrators can back up their entire virtual infrastructure in seconds and support other data management possibilities. The snapshot copies can then be backed up to tape and replicated to another facility with IBM System Storage N series with SnapMirror® or IBM System Storage N series with SnapVault®. VMs can be restored almost instantly, individual files can be quickly and easily recovered, and clones can be instantly provisioned for test and development environments. OSSV can be incorporated along with snapshot copies for a very robust solution.

LMU IP SAN architecture

This section describes the Loyola Marymount University (LMU) IP SAN architecture.

All ESX servers connect to an IBM System Storage N series N5500 via an iSCSI SAN. In order to utilize VMotion, the ESX servers are all in one initiator group (igroup).

Each VMFS volume is a LUN in the same flexible volume with its corresponding RDM LUNs. We limit the number of RDM LUNs in a single flexible volume because we want to limit the number of simultaneous VMware snapshots that the ESX hosts must handle. There is a maximum of 25 RDMs per flexible volume, and we store the .vmx files in a VMFS LUN in the same flexible volume. That allows us to group VMs according to the flexible volume to which they belong. A 50GB VMFS volume is large enough to handle the .vmx files, COW files and other configuration files used by ESX.

Although the VMFS LUN is included in the snapshot along with all the RDMs, there is no guarantee of consistency for VMFS in snapshots. This is not a significant concern because VMFS is used only to store .vmx files, RDM pointers, and VMware snapshot change logs (COW files). The loss of either of those would not be detrimental to the backup set.

Because .vmx files are small, you can use scp to send backup files to another location or install a traditional backup agent on an ESX host. The backup processes for these files should not add significant load in your environment.

Future plans for the LMU iSCSI SAN architecture include clustered head units and SnapMirror to replicate critical VMs at an alternate disaster recovery (DR) site. Here are several of the hardware configuration details that are architecturally compliant with the VMware and N series solution:

- IBM System Storage N series with Data ONTAP® 7.0.02 running on N5500
- iSCSI software initiators on ESX hosts (ESX servers are blade servers without TCP offload engine, or TOE, capability)

- ESX servers are limited to 32 snapshots for each single physical VM host[1]
- Cluster of 7 ESX Hosts, hosting 54 VMs (multiple disk I/Os / logically separated I/Os).

Backup scenarios

Figure 1 shows normal operations, and Figure 2 shows operations following snapshot restoration.

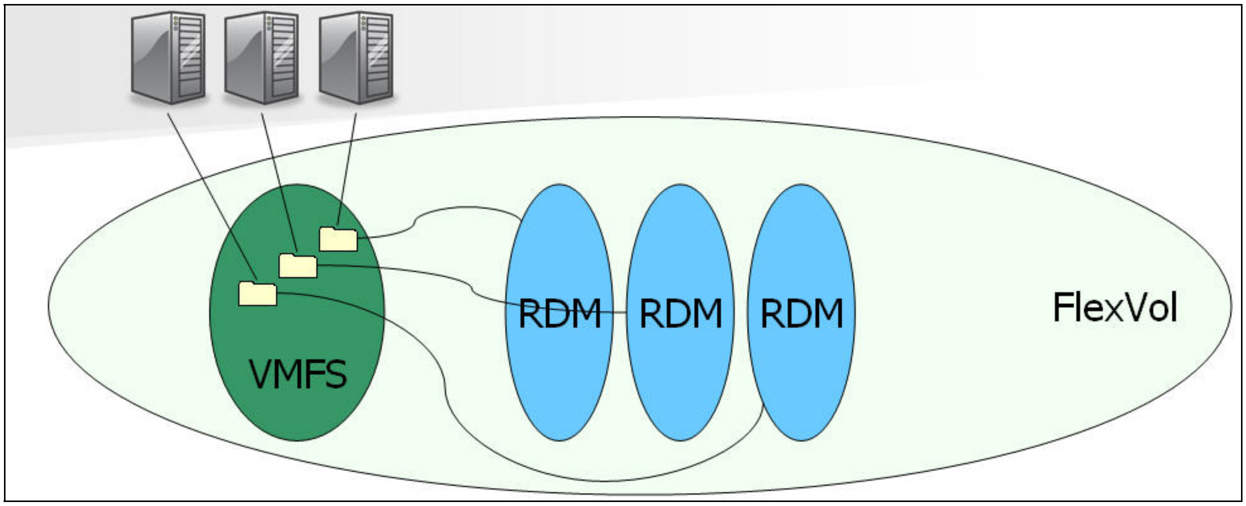

Figure 1) Normal Operations.

[1] This is a VMware snapshot specific limitation. N series can exceed this limitation and have up to 255 snapshots per volume.

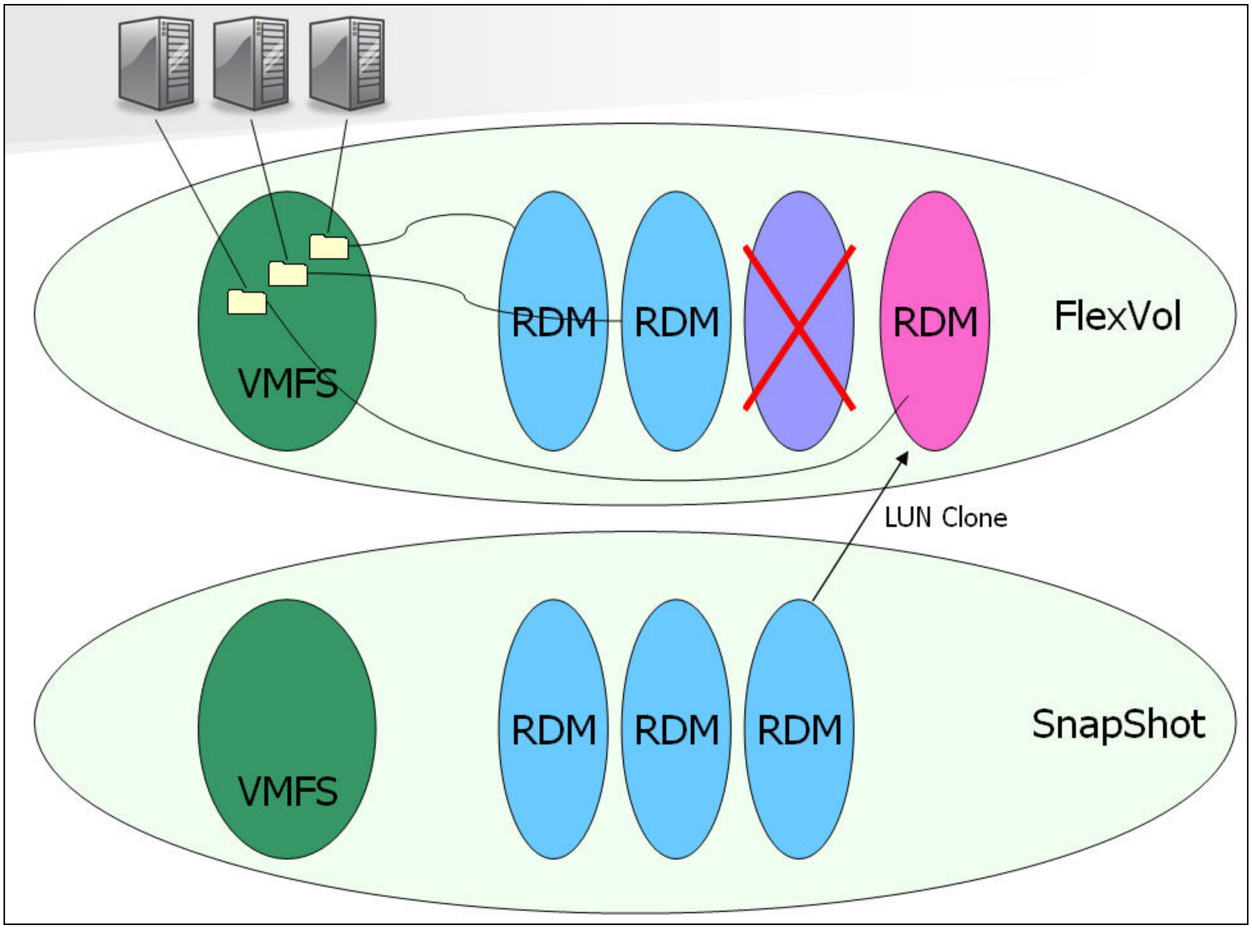

Figure 2) Operations following snapshot restoration.

The appendix contains a complete script written by Jason Guibert, Systems Administrator from LMU, which he designed to perform the following functions:

- Quiesce the RDM LUN
- Take the snapshot
- Unquiesce the RDM LUN.

The script leverages the features of several utilities: VCB, VMware Virtual Center (VMVC), and the osql utility. Because VMVC uses a SQL database to store information, we can access this data directly with the osql command line utility.

VCB comes with several command-line utilities; it is a set of scripts designed to plug in to backup software. The scripts all use the same core command line utilities. You can write your own scripts to use the VCB utilities for other purposes. The VCB utilities access VMVC's SQL database to get information about the target VM. In this example, we will use vcbsnapshot. This command locates the proper host for the VM and creates the VMware snapshot:

```
vcbsnapshot -h localhost -u username -p password -c moref:vm-611 quiesce**
```

In the LMU setup, VCB is installed on the same Windows server as VMVC (thus the local host). But the target hostname should be the VMVC server. The username and password must be an account with appropriate permissions in VMVC. You can store the username and password in a configuration file (config.js) so that it is not necessary to write them on the command line every time. You can also declare all environment-specific values in environment variables at the beginning of the scrip, such as VMVC hostname, usernames and passwords. The target VM is formatted “moref:vm-611,” which means “machine object reference,” and 611 is the record ID in SQL. “quiesce” is the name of the snapshot. Because command is sent to VMVC and VMVC knows on which host the VM resides, it automatically forwards the command to the appropriate ESX host.

To remove the snapshot created above, the following command is used:

```
vcbsnapshot -h localhost -u username -p password -d moref:vm-611 ssid:snapshot-619
```

It is necessary to know the snapshot ID. How are these ID numbers obtained? The answer is using osql. osql is a command-line utility that comes with SQL Enterprise Manager. You need only the osql.exe file, which you can copy from your installation of SQL. Because VMVC stores its data in a SQL database, you can use osql to query the VMVC database and get the necessary ID numbers for both the VMs and the snapshots:

```
osql -U username -P password -D VirtualCenter -Q "SELECT ID FROM vc_user.VPX_VM WHERE
IS_TEMPLATE = 0"
```

The command above connects to the database referenced by the System database source name (DSN), in your open database connectivity (ODBC) connectors, called “VirtualCenter.” It logs into the database with the listed username and password, which must have rights on the database in SQL. Then the actual SQL query pulls the record ID number for each VM that is not a template. (You can customize the filter criteria here. The full script in the appendix uses a query that isolates the VMs according to the location of their .vmx files). In the example, “vc_user” is the database owner in SQL. “vpx_vm” is the SQL table that stores the list of all VMs.

The output from the previous command looks like this:

```
ID
-----------
108
110
…
1226

(19 rows affected)
```

Similarly, to get the snapshot ID, we use:

```
osql -U username -P password -D VirtualCenter -Q "select VM_ID, ID from
vc_user.VPX_SNAPSHOT where SNAPSHOT_NAME = 'Quiesce'"
```

This query returns a tab-separated list of the VM ID and snapshot ID for each VM that has a snapshot named “quiesce.” The output looks like this:

```
VM_ID ID
----------- -----------
108 1228
110 1229
… …
```

```
1226 1246

(19 rows affected)
```

The next step is to configure the output of the command so that it can be used in a script. Add the –o switch to direct output to a text file. Add the switch “-h-1” to remove the column headers. Then, when you run the VCB snapshot command in a “FOR” loop, you set the “(“ character as the end-of-line (EOL) character. That will cause the FOR loop to skip the last line of your output that says “(## Rows Affected).”

```
osql -U vc_user -P vmware -D VirtualCenter –h-1 -Q "SELECT ID FROM vc_user.VPX_VM
WHERE IS_TEMPLATE = 0" –o output.txt
```

The last component is the set of commands that manage the snapshots. Remote shell (RSH) is used to connect to the N series storage system. You need a user account with RSH permissions in N series; it can be a local user account on the Windows box where the script is run. Modify /etc/hosts.equiv on N series to grant RSH permission or use the IBM System Storage N series with FilerView® GUI. The command is illustrated as follows:

```
RSH [nameofNseries] snap create volname snapshotname
```

Similarly, there are also “snap rename” and “snap delete” commands to manage old snapshots.

Summary

Automated VMware backups with N series Snapshot technology are a highly effective solution for host-based backups in a SAN environment. N series systems are the ideal storage platform for a virtual infrastructure and can address VMware challenges in a way that is unparalleled in the storage market.

This document described a best practices approach that has already been applied in a production environment. Note that not all factors are addressed, however, and that expertise is required to solve user-specific deployments.

Appendix: Backup script

This appendix contains a backup script to perform the operations described in this document.

```
-------- BEGIN SCRIPT --------
@echo off
echo Script Execution Commenced at %date% %time% >> snapall.log
REM Define machine names, usernames, and passwords
setlocal
set filer=replace this text with the DNS name of your N series filer
set volname=replace this text with the N series name of the target flexvol
set sqlpword=replace this text with the password for the database owner (DBO) in SQL
set dbdsn=replace this text with the DSN of your VMVC database from your ODBC data source
set vcenter=replace this text with the DNS name of your VMVC server
set vcuser=replace this text with a username on VNVC with rights to create and delete snapshots
set vcpword=replace this text with the password for the above user

REM change to directory containing VCB utilities, osql.exe, and this script
c:
cd "\Program Files\VMware\VMware Consolidated Backup Framework"

REM Manage old snapshots on N series
REM In this example scenario we retain 4 snapshots.
echo deleting vmware_snap4 >> snapall.log
RSH %filer% snap delete %volname% vmware_snap4 >> snapall.log
echo renaming vmware_snap3 to vmware_snap4 >> snapall.log
RSH %filer% snap rename %volname% vmware_snap3 vmware_snap4
echo renaming vmware_previous to vmware_snap3 >> snapall.log
RSH %filer% snap rename %volname% vmware_previous vmware_snap3
echo renaming vmware_recent to vmware_previous >> snapall.log
RSH %filer% snap rename %volname% vmware_recent vmware_previous

REM Get list of VMs from SQL database and output to vmlist.txt
REM "vc_user" is the example name of the SQL database owner.
Replace all occurrences of "vc_user" with the actual name of your DBO for the VMVC database in
SQL.
REM 688 is the example value for the datastore ID.
Look up the actual value in your VPX_DS_ASSIGNMENT table in SQL and replace it in the line
below.
REM (this example script assumes you only have one FlexVol to snapshot).
osql -U vc_user -P %sqlpword% -D %dbdsn% -h-1 -Q "SELECT vc_user.VPX_VM.ID FROM
vc_user.VPX_VM INNER JOIN vc_user.VPX_DS_ASSIGNMENT ON vc_user.VPX_VM.ID =
vc_user.VPX_DS_ASSIGNMENT.ENTITY_ID where ds_id = 688" -o vmlist.txt

REM quiesce all VMs by creating VMware snapshots with snapshot name = Quiesce.
for /F "eol=(" %%i in (vmlist.txt) do echo Creating snapshot for vmid=%%i >> snapall.log &&
vcbsnapshot -h %vcenter% -u %vc_user% -p %vcpword% -c moref:vm-%%i Quiesce >> snapall.log

REM create new snapshot on N series
echo creating vmware_recent >> snapall.log
RSH %filer% snap create %volname% vmware_recent >> snapall.log

REM Get list of VMware snapshot IDs from SQL database and output to sslist.txt
```

```
osql -U vc_user -P %sqlpword% -D %dbdsn% -h-1 -Q "select VM_ID,ID from
vc_user.VPX_SNAPSHOT where SNAPSHOT_NAME = 'Quiesce'" -o sslist.txt

REM unquiesce all VMs by removing all VMware snapshots.
for /F "eol=( tokens=1,2" %%i in (sslist.txt) do echo Removing snapshot for vmid=%%i >> snapall.log
&& vcbsnapshot -h %vcenter% -u %vcuser% -p %vcpword% -d moref:vm-%%i ssid:snapshot-%%j
>> snapall.log

if errorlevel 0 goto End
:Error
echo An error occurred during script execution >> snapall.log

:End
echo Script Execution complete at %date% %time% >> snapall.log

REM to run this script on other volumes, create a new instance of the script and change the volname
and ds_id.
-------- END SCRIPT --------
```

Trademarks and special notices

References in this document to IBM products or services do not imply that IBM intends to make them available in every country.

Data ONTAP, Network Appliance, the Network Appliance logo, SnapMirror and Snapshot are trademarks or registered trademarks of Network Appliance, Inc., in the U.S. and other countries.

Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

Other company, product, or service names may be trademarks or service marks of others.

Information is provided "AS IS" without warranty of any kind.

All customer examples described are presented as illustrations of how those customers have used IBM products and the results they may have achieved. Actual environmental costs and performance characteristics may vary by customer.

Information concerning non-IBM products was obtained from a supplier of these products, published announcement material, or other publicly available sources and does not constitute an endorsement of such products by IBM. Sources for non-IBM list prices and performance numbers are taken from publicly available information, including vendor announcements and vendor worldwide homepages. IBM has not tested these products and cannot confirm the accuracy of performance, capability, or any other claims related to non-IBM products. Questions on the capability of non-IBM products should be addressed to the supplier of those products.

Performance is based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput or performance that any user will experience will vary depending upon considerations such as the amount of multiprogramming in the user's job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve throughput or performance improvements equivalent to the ratios stated here.

Any references in this information to non-IBM Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this IBM product and use of those Web sites is at your own risk.