VMware Update Manager Performance and Best Practices

VMware Update Manager 1.0

VMware Update Manager (VUM) provides a patch management framework for VMware Virtual Infrastructure. IT administrators can use it to patch VMware ESX, Windows, and certain versions of Linux virtual machines.

As data centers get bigger, performance implications become more important for patch management. This study covers the following topics:

Benchmarking Methodology

Experimental Setup

VUM 1.0, VMware VirtualCenter 3.5, and ESX 3.5 were used for performance measurements. WANem 1.2 was used for simulating a high-latency network with packet drops. Microsoft Windows XP SP2 was used for powered-off virtual machine scan and remediation. Red Hat Enterprise Linux 32-bit was used for Linux virtual machine scan.

VUM and VirtualCenter Server

Host Computer: Dell PowerEdge 2970

CPUs: Two 2GHz AMD Opteron 2212 dual-core processors

RAM: 16GB

Hard drives: Eight 73GB SAS drives

Network: Broadcom NetXtreme II5708 1Gbps

VUM Server software: VUM 1.0

VirtualCenter Server software: VMware VirtualCenter 3.5

ESX System

Host Computer: Dell PowerEdge 2900

CPUs: Two 2.66GHz Intel Xeon 5355 quad-core processors

RAM: 32GB

Hard drives: Eight 73GB SAS drives

Network: Broadcom NetXtreme II5708 1Gbps

ESX software: VMware ESX 3.5

Virtual Machine Operating Systems

Windows: Microsoft Windows XP SP2

Linux: Red Hat Enterprise Linux 32-bit (kernel: 2.6.18-8.el5)

Network Simulation Software

WANem 1.2

Network Configurations

The network configurations used in the experiments are shown in Figure 1 (basic network configuration) and Figure 2 (high-latency network configuration).

Figure 1. Basic Network Configuration

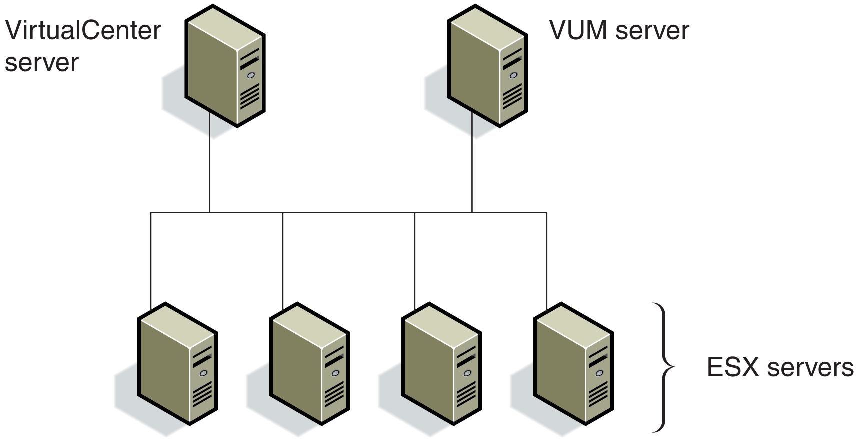
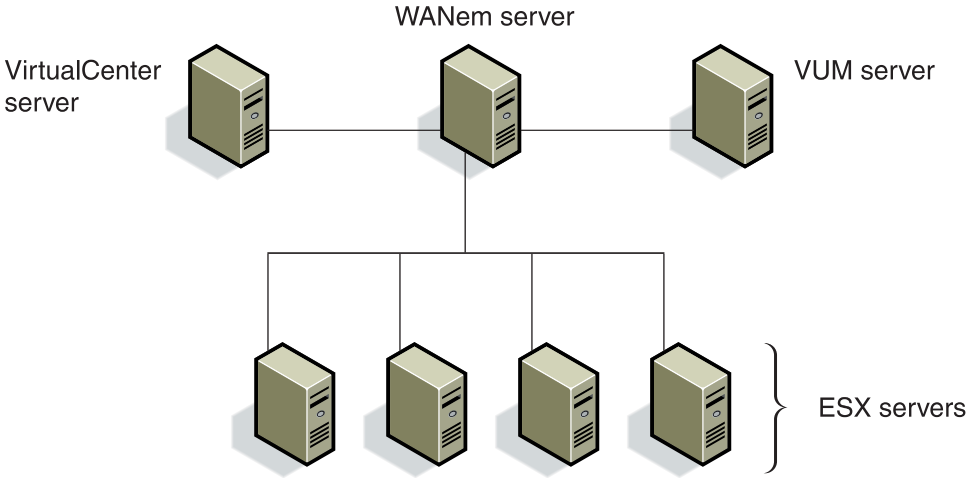


Figure 2. High Latency Network Configuration



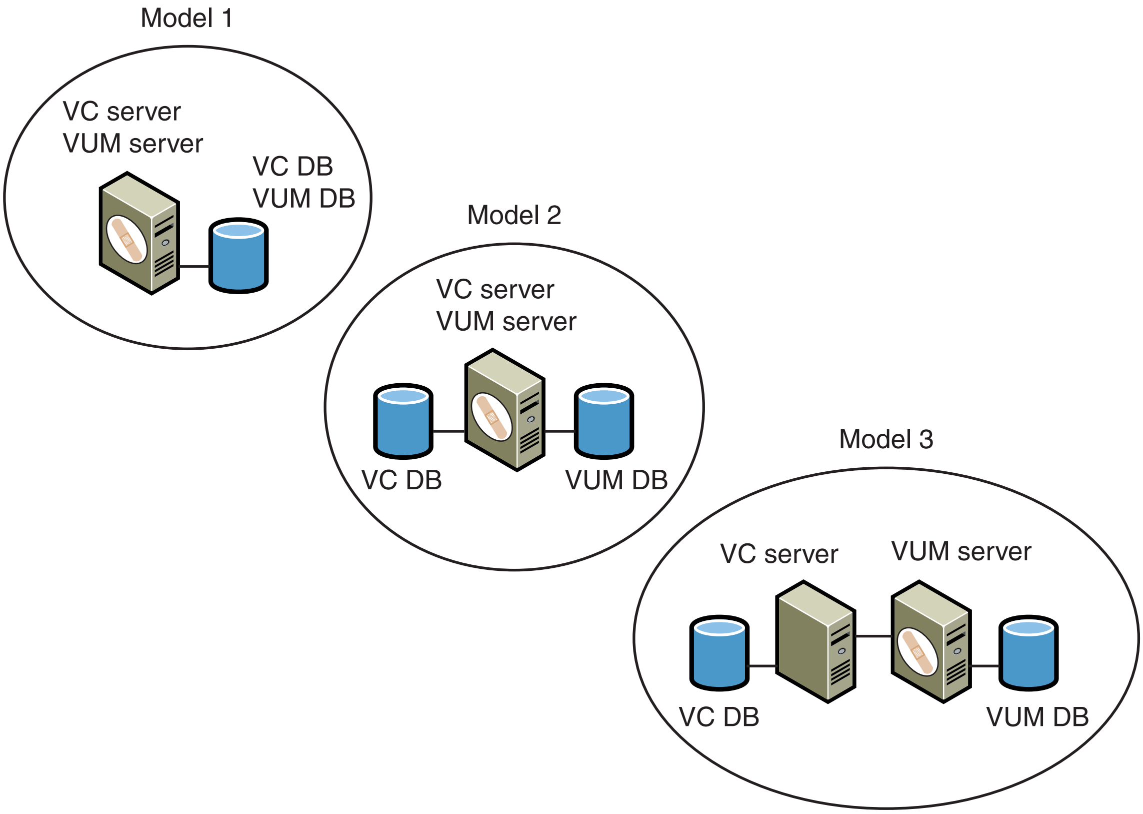
VUM Server Host Deployment

The three VUM server host deployment models are shown in Figure 3. In model 1 VirtualCenter and the VUM server share both a host and a database instance. In model 2, recommended for datacenters with more than 500 virtual machines or 50 ESX hosts, VirtualCenter and the VUM server still share a host, but use separate database instances. In model 3, recommended for datacenters with more than 1,000 virtual machines or 100 ESX hosts, VirtualCenter and the VUM server run on different hosts, each with its own database instance.

For both ESX and virtual machine patching the VUM server transfers patch files over the network. To avoid unnecessary disk I/O it is ideal if the VUM server host can cache patch files, some of which are several hundred megabytes, completely within system cache. To this end it is desirable for the VUM server host to have at least 2GB of RAM.

It is also best to place the patch store and VUM database on separate physical disks. This arrangement distributes the VUM I/O and dramatically improves performance.

Figure 3. VUM Server Host Deployment Models



Performance Tips

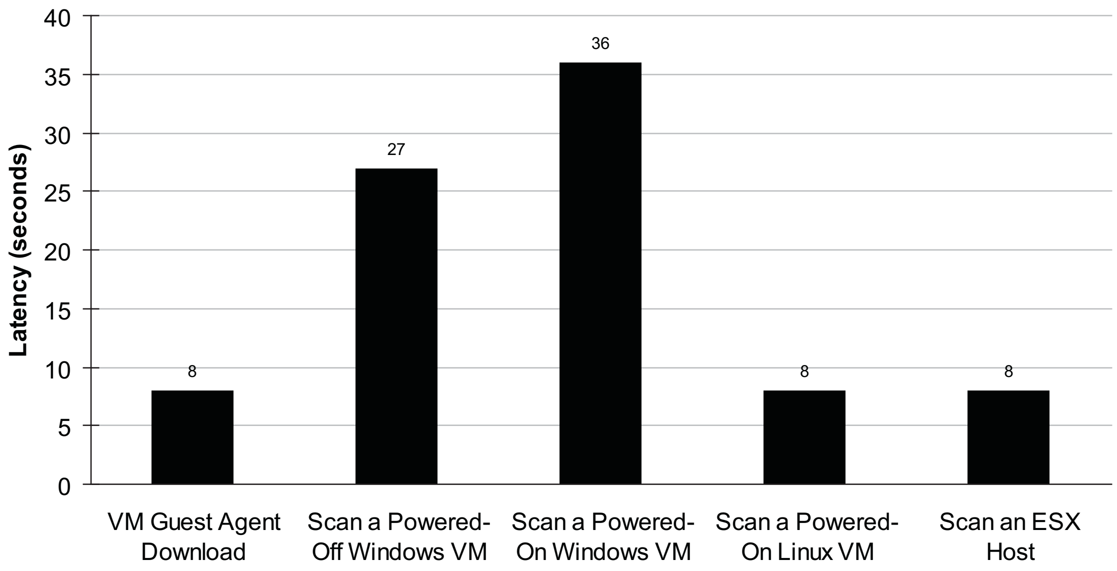
- Separate the VUM database from the VirtualCenter database when there are 500+ virtual machines or 50+ hosts.
- Separate both the VUM server and the VUM database from the VirtualCenter server and the VirtualCenter database when there are 1000+ virtual machines or 100+ hosts.
- Make sure the VUM server host has at least 2GB of RAM to cache patch files in memory.
- Allocate separate physical disks for the VUM patch store and the VUM database.

Latency Overview

VUM operation latency is an important metric. IT administrators need to finish applying patches within maintenance windows. Figure 4 shows latency results for guest agent download, powered-off Windows virtual machine scan, powered-on Windows virtual machine scan, powered-on Linux virtual machine scan, and ESX host scan.

Guest agent download consists of multiple steps. The number presented in Figure 4 is the time required to send the file through the network from the VUM server to the virtual machine. For both the powered-on and powered-off virtual machine scan performance data presented in Figure 4 the guest agent is already installed. All the numbers are averaged over multiple runs. The powered-on virtual machine scan showed higher deviation than did the powered-off virtual machine scan. Note that these latency numbers are only references. Actual latency varies significantly with different deployment setups.

Figure 4. VUM Operation Latency Overview



The results in Figure 4 were measured on a low-latency local network setup. However network latency plays an important role for most VUM operations. For example, Figure 4 shows a powered-off virtual machine scan beating a powered-on virtual machine scan. On a high-latency network, however, a powered-on virtual machine scan can perform better than a powered-off virtual machine scan because the powered-on scan transfers less data. For results obtained on a high-latency network see "Network Latencies" on page 8.

Performance Tips

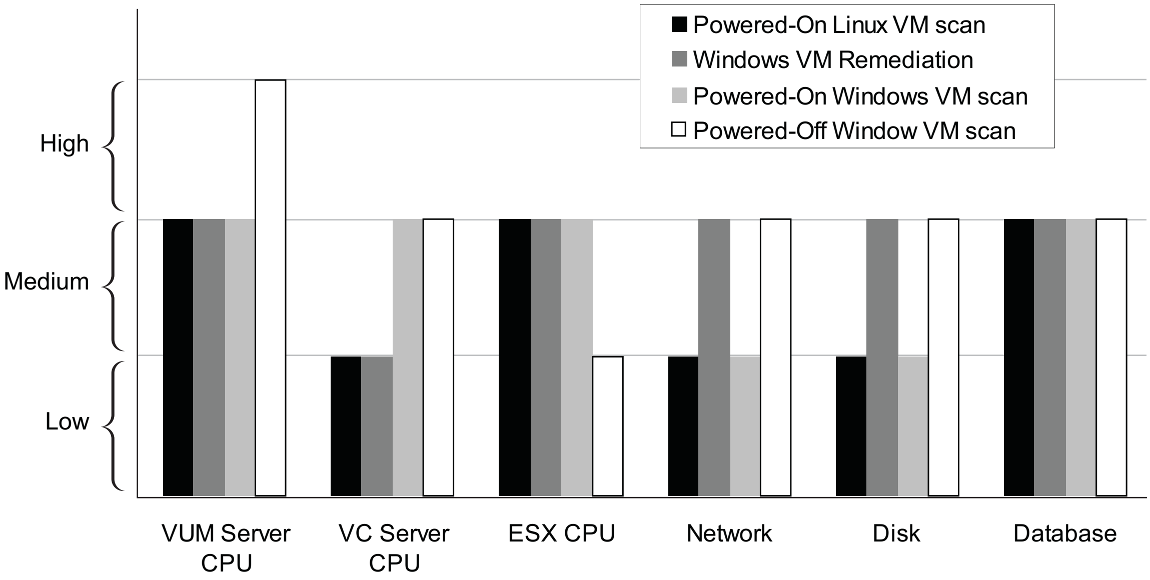
- Because the Windows guest agent is installed in each virtual machine the first time a powered-on scan is run, the first powered-on scan command can take longer than subsequent scans. It may therefore be desirable to run the first scan command when this additional time will not be an issue.

NOTE Because the Linux virtual machine guest agent is installed when a powered-on virtual machine is first added to the VMware Infrastructure inventory (rather than the first time a scan is run), this performance tip does not apply to Linux virtual machines.

Resource Consumption Matrix

VUM operations have different loads on the VUM server, VirtualCenter server, and ESX host. Figure 5 divides VUM operations into low, medium, and high resource consumption levels. For example, scanning a powered off Windows virtual machine results in high VUM server CPU usage, medium VirtualCenter server CPU, network, disk, and database usage, and low ESX CPU usage.

Figure 5. Resource Consumption for VUM Operations



To alleviate resource pressure on VUM servers and ESX hosts, VUM performs job throttling by limiting the maximum number of concurrent operations. Table 1 gives the default performance limits.

Table 1. Job Throttling Default Limits for Each VUM Operation

VUM Operations	Maximum Tasks per ESX Host	Maximum Tasks per VUM Server
VM Remediation	5	48
Powered-on Windows VM Scan	6	72
Powered-off Windows VM Scan	6	10
Powered-on Linux VM Scan	6	72
ESX Host Scan	1	72
ESX Host Remediation	1	48

The VUM server combines all operation types when calculating the limit. If a combination of operation types are being performed simultaneously, the formulas should be used with caution. For example, if powered-on Windows virtual machine scans and powered-off Windows virtual machine scans are being performed simultaneously, the maximum number of concurrent scans is *not* 12 (6+6) per ESX host, but 6 per ESX host.

Performance Tips

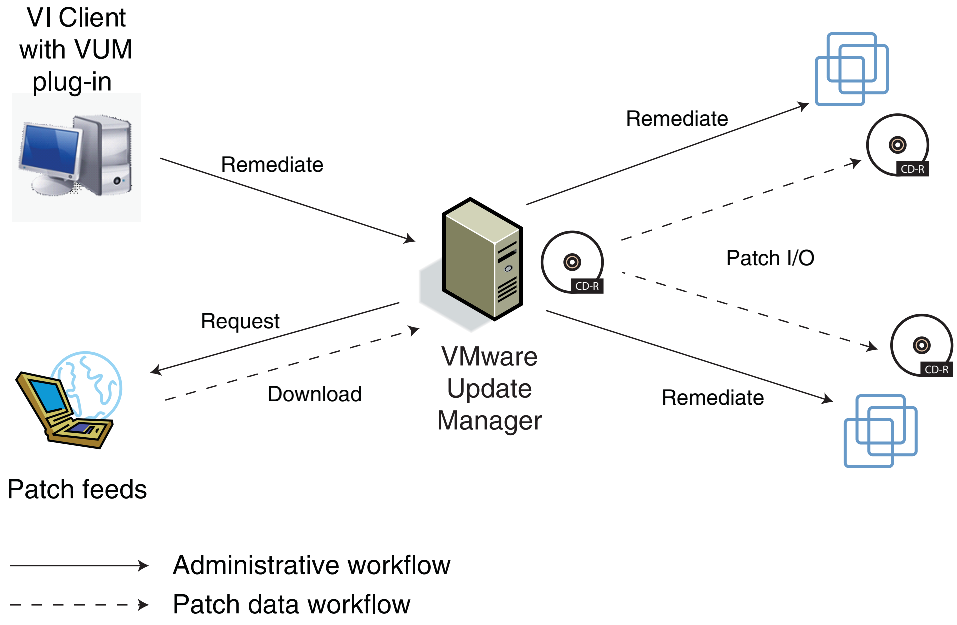
- For a large setup, powered-on virtual machine scans are preferred if VUM server resources are constrained or more concurrency is needed for scans.

Guest Operating System Tuning

The VUM guest agent installed on a virtual machine is single threaded for both powered-on virtual machine scans and virtual machine remediation. Multiple vCPUs thus do not reduce VUM operation latency.

The amount of RAM available to the guest operating system, however, can have a significant effect on the performance of virtual machine remediation operations. As shown in Figure 6, when a remediation command is issued from the VI client through the VUM plug-in, the VUM server first downloads the patches, creates an ISO patch file, and mounts it as a virtual CDROM to the guest operating system. The guest agent inside the guest operating system verifies the checksum of the patch file before the actual patch is applied. If the guest operating system RAM size is too small to cache the entire patch file into system cache, the patch file has to be read again. Because some Windows service packs can be as big as a few hundred megabytes in size, it is best to configure each virtual machine with at least 1GB of RAM so duplicate reads for the whole patch files can be avoided.

Figure 6. Guest Operating System Patching Flow



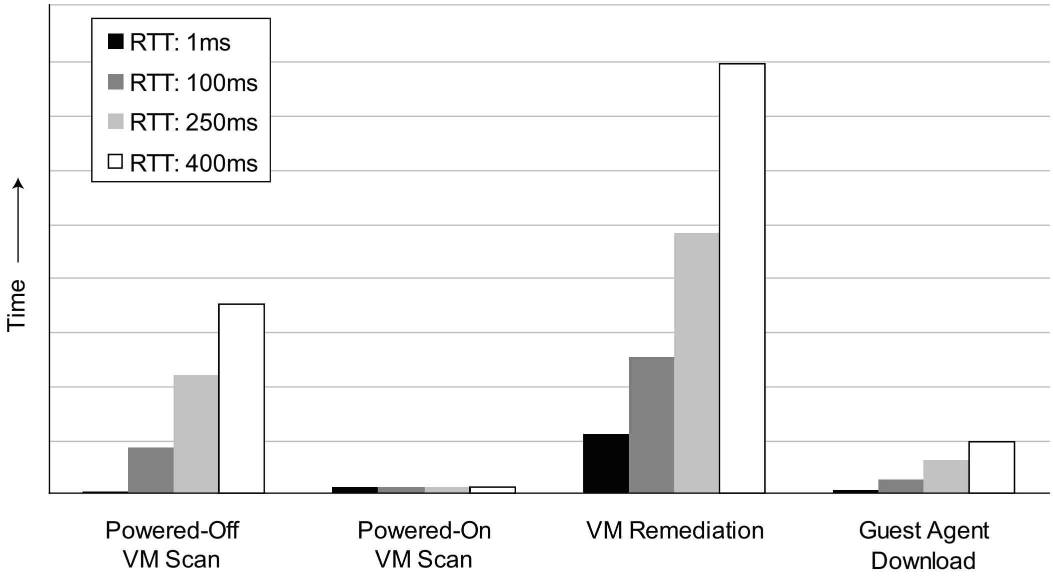
Performance Tips

- Multiple vCPUs do not help VUM operations as the VUM guest agent is single threaded.
- Configure each virtual machine with at least 1GB of RAM so large patch files can fit in the system cache.

Network Latencies

VUM server performance is strongly affected by network speeds. For powered-off virtual machine scans, the VUM server needs to read the registry information remotely from the ESX host. For virtual machine remediation, the VUM guest agent reads the patch ISO file through the network from the VUM host. The latency results for VUM operations would thus be expected to be impacted when the network latency increases. Figure 7 shows the latency results for powered-off virtual machine scans, powered-on virtual machine scans, virtual machine remediation, and guest agent download. As shown in Figure 2, WANem is used in this test to simulate a high latency network.

Figure 7. Operation Latency with a High Latency Network



Network round trip times (RTTs) are typically 75 to 100 milliseconds for intra-U.S. networks, about 250 milliseconds for transatlantic networks, and 320 to 430 milliseconds for satellite networks.

In Figure 7, the latency of powered-off virtual machine scans, virtual machine remediation, and guest agent download operations increases linearly with the network speed. Powered-on virtual machine scans stay flat as most of the job is done inside the virtual machine and only the scan result info is transferred over the network.

For an unreliable network with packet drops, both powered-off virtual machine scans and virtual machine remediation are impacted significantly. The root cause is that the TCP retransmission timer varies from a few hundred milliseconds to a few seconds depending on the latency of the network. With a high-latency network, the TCP retransmission timer is much longer than with a low-latency network. It is therefore highly desirable to deploy the VUM server close to the ESX hosts to avoid network latency and packet drop.

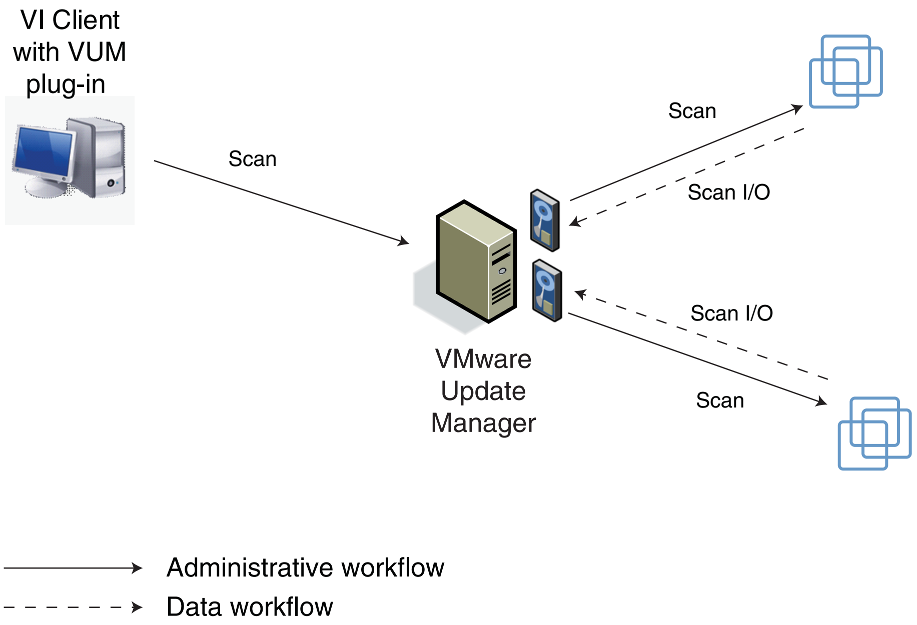
Performance Tips

- Deploy the VUM server close to the ESX hosts if possible. This reduces network latency and packet drops.
- On a high-latency network, powered-on virtual machine scans are preferred as they are not sensitive to network latency.

On-Access Virus Scanning

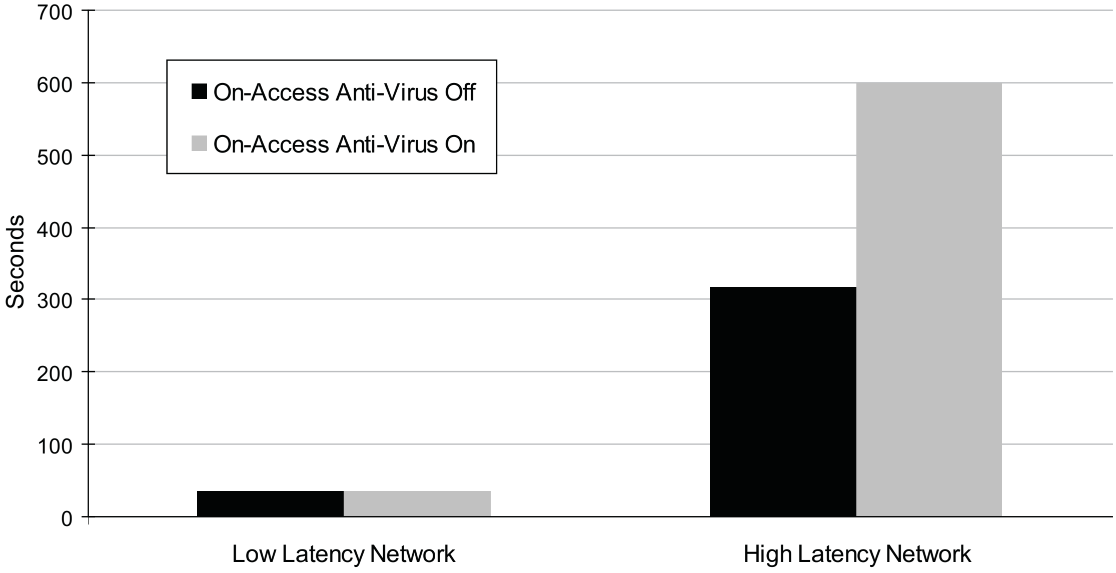
As shown in Figure 8, for powered-off or suspended virtual machines a scan command is issued from the VI client with the VUM plug-in and the VUM server mounts the virtual machine as a disk in the VUM server host. The VUM server then copies the hive files of the virtual machine and scans each file in the registry one by one. The file scan process has been optimized to read only necessary blocks of the scanned files to determine the version information. On average, there are only four reads per scanned file regardless of the actual size of each file. This is very important because the network between the VUM server and ESX host might be a high-latency network and reducing traffic on such a network can improve the performance dramatically.

Figure 8. Scan for Powered-Off or Suspended Virtual Machine



With on-access virus scanning, however, the optimization for reading only necessary blocks has no effect. When a file is opened, the on-access virus scanner will try to read the whole file to scan for viruses. This means all blocks of a scanned file will be fetched through the network. This is less of an issue on a low-latency network for powered-off scans as the file transfer speed is very high and thus the time to transfer the files can be ignored. Figure 9 show that on a low-latency network the total time to do powered-off scans is about the same with on-access virus scanning on or off. On a high-latency network, however, turning on-access virus scanning off can improve powered-off scan latency by almost 50%.

Figure 9. Powered-off Scan Latency With McAfee On-Access Anti-Virus



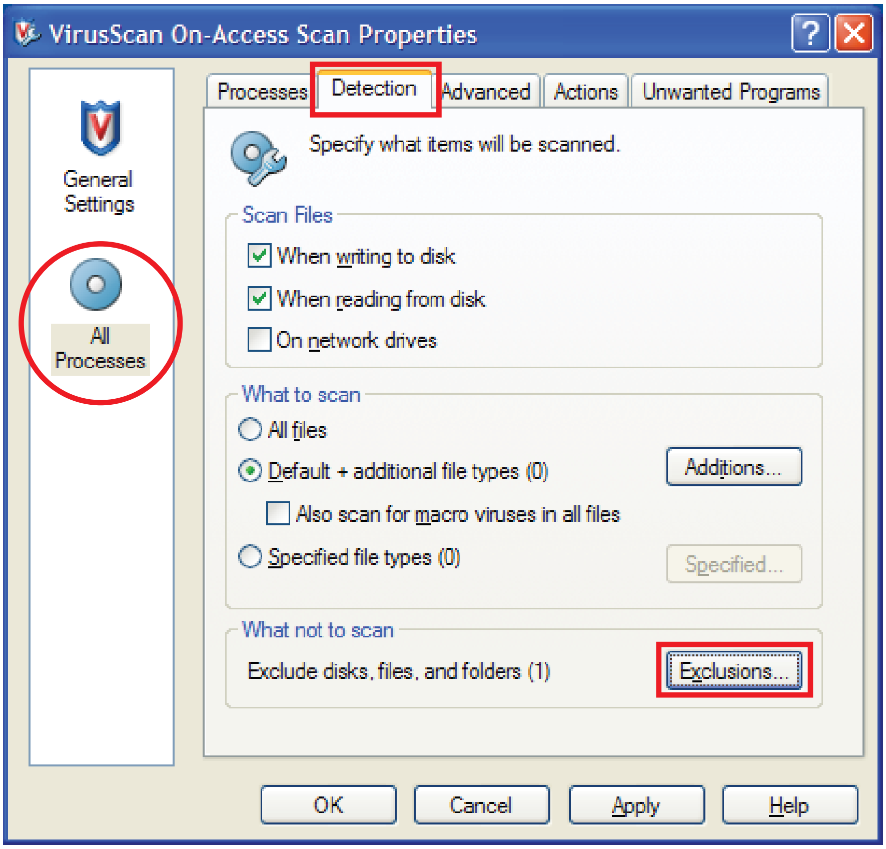
It is recommended that the mounted disk be excluded from on-access virus scanning to achieve optimal performance. Here is an example of how to exclude it for McAfee On-Access Anti-Virus.

1 Click the **All Processes** icon on the left of the **VirusScan On-Access Scan Properties** window (see Figure 10).

2 Select the **Detection** tab.

3 Click the **Exclusions...** button.

Figure 10. VirusScan On-Access Scan Properties Window



Specify what items will be scanned.

Scan Files

☑ When writing to disk

☑ When reading from disk

☐ On network drives

What to scan

○ All files

● Default + additional file types (0)

☐ Also scan for macro viruses in all files

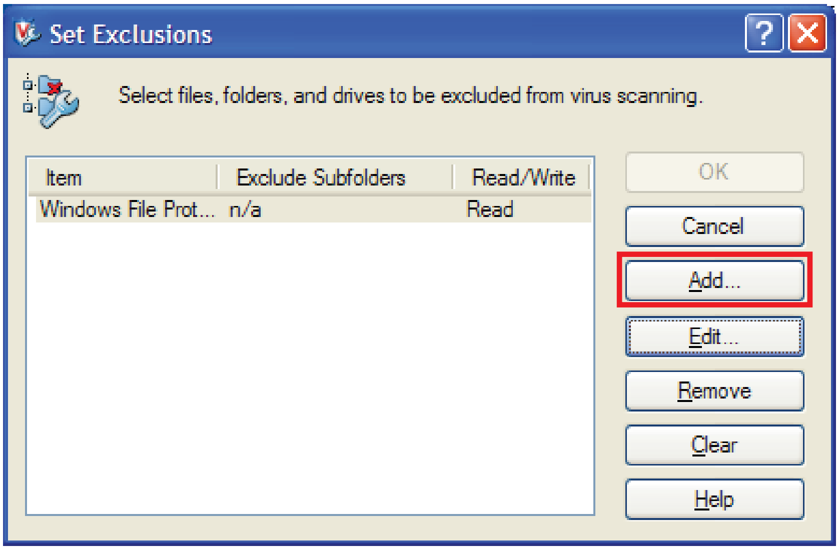
○ Specified file types (0)

What not to scan

Exclude disks, files, and folders (1)

4 Click the **Add** button (see Figure 11).

Figure 11. VirusScan Set Exclusions Window



Set Exclusions

Select files, folders, and drives to be excluded from virus scanning.

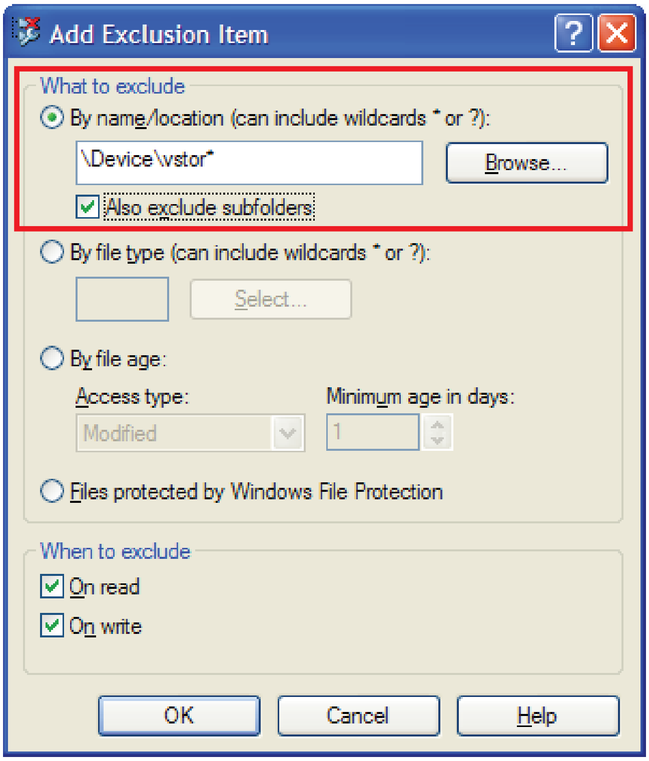
Item	Exclude Subfolders	Read/Write
Windows File Prot...	n/a	Read

5 Select the **By name/location** radio button (see Figure 12).

6 Type:
\Device\vstor*

7 Select the **Also exclude subfolders**, **On read**, and **On write** checkboxes.

Figure 12. VirusScan Add Exclusion Item Window



Add Exclusion Item

What to exclude

● By name/location (can include wildcards * or ?):

\Device\vstor*

☑ Also exclude subfolders

○ By file type (can include wildcards * or ?):

○ By file age:

Access type: Modified

Minimum age in days: 1

○ Files protected by Windows File Protection

When to exclude

☑ On read

☑ On write

8 Click **OK** in the **Add Exclusion Item** window.

9 Click **OK** in the **Set Exclusions** window.

10 Click **OK** in the **VirusScan On-Access Scan Properties** window.

Performance Tips

- Check if on-access virus scanning software is running on the VUM server host. If it is, exclude the mounted disk, as described in this section.

Conclusion

VMware Update Manager delivers the most full-featured and robust patch management product for VMware Infrastructure. In this white paper, the following performance recommendations have been made:

- Separate the VUM database from the VirtualCenter database when there are 500+ virtual machines or 50+ hosts.
- Separate both the VUM server and the VUM database from the VirtualCenter server and the VirtualCenter database when there are 1000+ virtual machines or 100+ hosts.
- Make sure the VUM server host has at least 2GB of RAM to cache patch files in memory.
- Allocate separate physical disks for the VUM patch store and the VUM database.
- Because the Windows guest agent is installed in each virtual machine the first time a powered-on scan is run, the first powered-on scan command can take longer than subsequent scans. It may therefore be desirable to run the first scan command when this additional time will not be an issue.
- For a large setup, powered-on virtual machine scan is preferred if VUM server resources are constrained or more concurrency is needed for scans.
- Multiple vCPUs do not help VUM operations as the VUM guest agent is single threaded.
- Configure each virtual machine with at least 1GB of RAM so large patch files can fit in the system cache.
- Deploy the VUM server close to the ESX hosts if possible. This reduces network latency and packet drops.
- On a high-latency network, powered-on virtual machine scans are preferred as they are not sensitive to network latency.
- Check if on-access virus scanning software is running on the VUM server host. If it is, exclude the mounted disk on a high-latency network.

References

- Administration Guide - Update Manager
  http://www.vmware.com/pdf/vi3_vum_10_admin_guide.pdf
- VMware Update Manager 1.0 Sizing Estimator
  http://www.vmware.com/support/vi3/doc/vi3_vum_10_sizing_estimator.xls

If you have comments about this documentation, submit your feedback to: docfeedback@vmware.com

VMware, Inc. 3401 Hillview Ave., Palo Alto, CA 94304 www.vmware.com

Item: EN-000157-00