VMware® Horizon View™ 5.2 Performance and Best Practices

Performance Study

TECHNICAL WHITE PAPER

Table of Contents

Introduction

VMware® Horizon View™ 5.2 liberates the end user identity from being locked to a particular system and enables anywhere computing. View transforms IT by simplifying and automating desktop management. IT administrators can compose virtual desktops on demand, based on locations and profiles, and can deliver at any time to desktops and devices anywhere. As shown in Figure 1, by centrally maintaining desktops, applications, and data, VMware Horizon View reduces costs, improves security and, at the same time, increases availability and flexibility for end users.

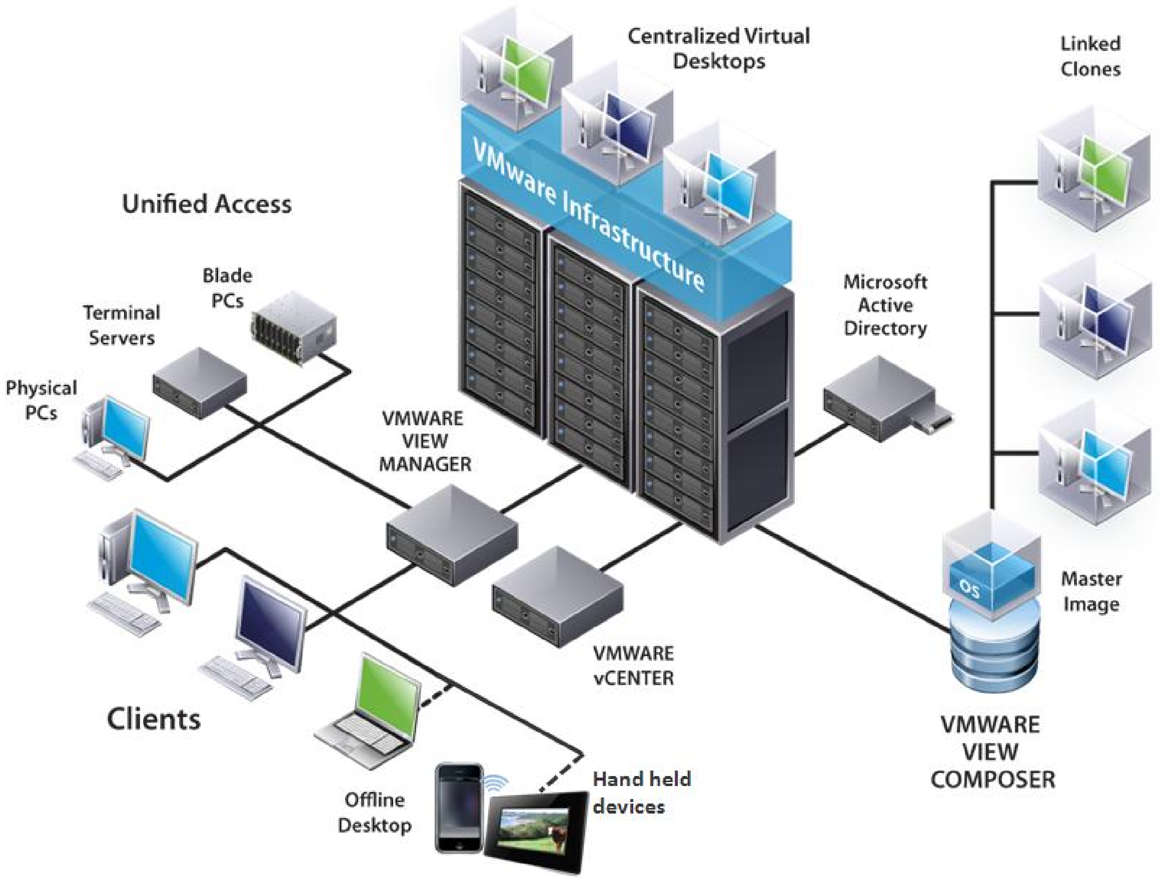

Figure 1. View architecture diagram

VMware Horizon View 5.2 offers the best end-user experience and performance over LAN, WAN, and extreme WAN. The VMware View PCoIP display protocol's adaptive technology provides an optimized virtual desktop delivery on both LAN and WAN. View 5.2 addresses the broadest list of use cases and deployment options with a single protocol and enables access to personalized virtual desktops complete with applications, end-user data, and settings anywhere, at any time. View 5.2 also reduces the overall costs of desktop computing by centralizing management, administration, and resources, and by removing IT infrastructure from remote offices.

View 5.2 New Features and Performance Improvements

This section briefly describes View 5.2 new features such as access of View desktops with Horizon, space efficient sparse (SEsparse) disks, hardware accelerated 3D graphics, and full support of Windows 8 desktops. View 5.2 performance improvements in the PCoIP protocol and View management are highlighted.

Accessing View Desktops with Horizon

With the Integration of View with Horizon, end users can now access their entitled View desktops through the same Horizon portal that they used to access their applications. Now the Horizon Web interface provides a single point of access for providing View desktops, user data, and applications. Whenever a View administrator includes a user or a group in a desktop pool, this entitlement is synced to Horizon and the desktops are made available to that user or the group from the Horizon interface. End users can utilize single sign on (SSO) and all the entitled resources including applications, desktops, and data are readily available using a single access point.

Support of Space Efficient Sparse (SESparse) Disks

In VMware vSphere® 5.1, a new disk format for VMs on VMFS was introduced to reduce the storage footprint of VMs. VMware Horizon View 5.2 leverages this disk format, also known as Space Efficient Sparse (SE sparse) disks, or elastic disks and provides the ability to reclaim those blocks that are unused or deleted by the guest file system and regain some of the space efficiencies promised by linked clones. This feature reduces storage costs and administrative overhead by efficiently using and reclaiming storage space to minimize the storage capacity requirements for persistent desktops and decreases the need to continuously recompose and restore images.

Support of Hardware Accelerated 3D Graphics

In response to user demand for an ever richer set of applications to be supported in the virtual environment, VMware has enhanced View to support hardware accelerated 3D graphics.

Support for non-hardware accelerated 3D graphics was introduced in VMware vSphere 5.0. This enabled VMware View 5.0 to support VMs running Windows Aero and enabled basic 3D application use (for example, Google Earth).

In the next phase of VMware's 3D vision, vSphere 5.1 introduced GPU virtualization, enabling multiple VMs to simultaneously share a single, physical GPU. This feature, termed vSGA (Virtual Shared Graphics Acceleration), is compatible with all key VMware technologies, including vMotion, and enables the recently released View 5.2 to support GPU-backed virtual desktops.

The support for hardware accelerated 3D graphics delivers significantly higher performance in a 3D environment. This expands the 3D application space that can be successfully run in View 5.2 to include key technologies such as CAD and medical imaging.

Support for 3D desktops in View can be administered by using the VMware Horizon View Administrator console, and can be enabled on a per-pool basis or controlled on a per-VM basis using the VMware vSphere client.

Optimizing Resource Sharing

In contrast to a physical workstation that has sole use of its GPU, in the virtualized environment GPUs become a shared resource. As a result, it is important to ensure that each VM does not use the GPU resource in a wasteful manner. For instance, in many situations it often does not make sense for a 3D application to render hundreds of frames per second if View is configured to remote at a lower frame-rate (30fps is the default setting). For these situations, View provides a registry setting to limit the maximum application frame rate. This can either be configured in the template VM or on a per-VM basis, and the value should typically be set to the maximum frame rate that is being used by PCoIP. This configuration is achieved by using the following registry setting (REG_DWORD):

```
HKLM\SOFTWARE\VMware, Inc.\VMware SVGA DevTap\MaxAppFrameRate
```

Setting this registry entry for a 3D workload has been found to significantly improve the performance and consolidation ratios achievable.

Windows 8 Support

In View 5.1, Microsoft Windows 8 support was offered as a tech preview, but with View 5.2, Windows 8 desktop VMs are now fully supported. Also with View 5.2, View clients are now supported in Windows 8 and users can utilize the PCoIP protocol to connect to the Window 8 desktop. Users can get RDP 8 protocol (when connected from a client with RDP 8 support); however, RDP 8 with the View client is not officially supported.

PCoIP Image Caching Improvements

In this release, PCoIP's image cache has been significantly improved to allow users on memory-constrained devices to run with much smaller cache sizes:

- **Enhanced scrolling:** Support was introduced to efficiently handle situations where image content is shifted vertically, as occurs during scrolling.
- **Improved cache compression:** View 5.2 debuts improved cache compression algorithms that provide significant additional compression of the View client's image cache.
- **Improved cache handling:** The cache's handling of progressive build operations has been made significantly more efficient.

All of these enhancements combine to allow users to derive significant bandwidth reductions using considerably smaller cache sizes than could be achieved with View 5.1.

Other PCoIP Improvements

Some other PCoIP improvements include:

- **GPO settings take immediate effect:** Many of the performance-oriented GPO settings now take effect immediately, allowing users or administrators to closely customize the behavior of their PCoIP sessions.
- **Relative mouse support:** Previously, support was only provided for absolute mode. However, for certain 3D applications, relative mouse is required and support is introduced on View 5.2.

View Management Improvements

There have been several significant enhancements for View management, including better UI response time for large pools, support of cluster with 32 hosts, and improved performance of View management operations. View 5.2 desktop architecture is simplified with a single VMware vCenter Server supporting up to 10,000 desktops in a pod. With support for 32 hosts per pool on VMFS along with NFS and pools spanning multiple VLANs, larger desktop pools can be created to decrease operational costs. The operational time of View management operations such as provisioning, recomposing, and rebalancing has improved significantly (by up to 2x) in View 5.2.

Performance Results

This section presents the View 5.2 PCoIP protocol performance results, Windows 8 and RDP 8 performance analysis, and the vSGA performance analysis, including how vSGA compares to the software renderer support introduced in View 5.1.

Experimental Setup

The experimental setup shown in Figure 2 was used for all the single VM protocol experiments. A Dell T610 tower server with a 2.53GHz Intel Xeon 5540 processor and 48GB physical memory hosted the desktop VM. The desktop VM's data and OS disk resided on a local SSD disk and it ran Windows 7 32-bit with 1 virtual CPU (vCPU) and 1GB of virtual memory. The desktop screen resolution is set to 1152 x 864 with true-color mode. A Windows XP 32-bit client is used to connect to the desktop VM with the selected remote display protocol and the client was configured with 1 vCPU and 768MB of virtual memory.

For the PCoIP protocol, build to lossless (BTL) is disabled using the GPO setting. Even though this setting is disabled, a typical office user gets perceptually lossless image quality. The maximum frame rate is set to 24 for every display protocol in all the experiments. ClearType text is also enabled to get clear and smooth fonts in all the experiments.

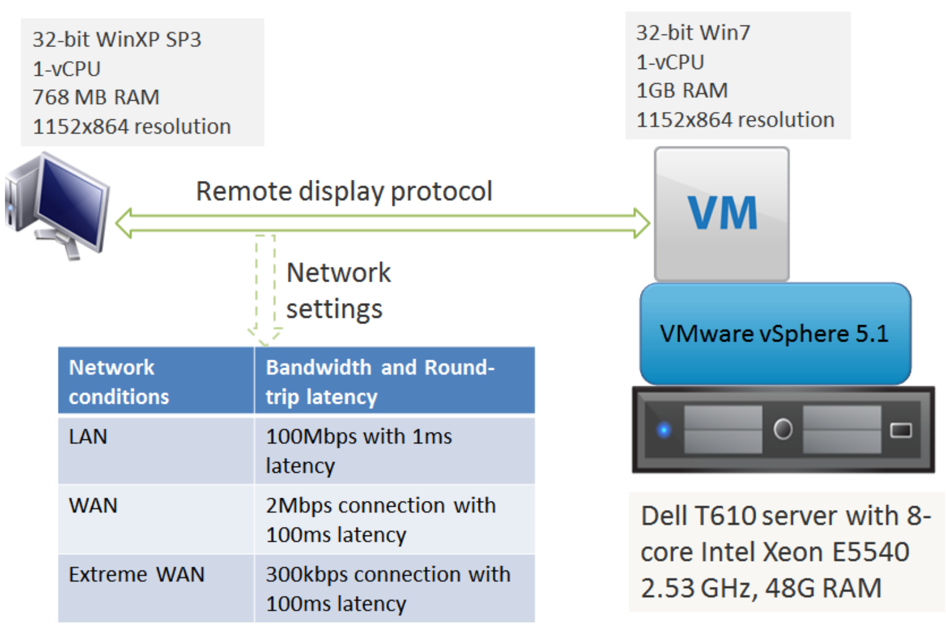

Figure 2. Experimental setup for comparing remote display protocol performance for different configurations. To study different user scenarios, different network conditions (LAN, WAN, extreme WAN) are being used.

Since the remote desktop can be accessed from a main office, branch office, or home office, the network conditions will be significantly different in all these cases. Therefore, three typical network conditions were used as shown in the experimental setup diagram.

- LAN – Plentiful bandwidth available with almost no latency.
- WAN – Bandwidth is limited (2 Mbps) and there is 100ms round-trip latency.
- Extreme WAN – Bandwidth is severely limited (300 kbps) and there is 100ms round-trip latency.

Workload

For all the performance experiments, we have used VMware View Planner, a workload generator that simulates typical office user operations, such as typing a word in MS Word, playing a PowerPoint slideshow, reading Outlook emails, browsing PDF pages, browsing Web pages, watching video, and so on. More details about the View Planner are described in the VMware View Planner user guide [1] and a recent article in *VMware Technical Journal* [2].

Figure 3. Applications used in View Planner workload. There is also a notion of "think time" which mimics real user behavior where users think for some time during their regular work.

As shown in Figure 3, View Planner supports different Office applications (and also supports Office 2010), other typical apps such as Adobe Reader, Microsoft Internet Explorer, Mozilla Firefox, and Microsoft Windows Media Player. All the applications were selected with medium video and a think time of 10 seconds was selected in all the experiments.

Also, for almost all experiments, the VMware View Planner workload was run for three network conditions (LAN, WAN, and extreme WAN) and the chart data were normalized to the "maximum" values seen in all network conditions. For example, if View 5.1 with a LAN configuration had the highest bandwidth usage, then it is normalized to 1.0 and other values are scaled with respect to that. In this example, View 5.1 with LAN is used as the baseline for comparison.

Using VMware View Planner, View 5.2 PCoIP performance is evaluated and the bandwidth usage results are presented (how many users can be supported on a given hardware platform and a network link) and host consolidation results (how many users can be consolidated on a single server) which is presented next.

Bandwidth Usage

To maximize user experience at constrained bandwidth environments, it is imperative for a display protocol to intelligently transfer less data while maintaining acceptable image quality. Hence, bandwidth usage for a Microsoft Office user workload is very important to know in order to estimate how many users can be supported on a given network link. This section focuses on the downlink bandwidth usage metric and presents related results.

In View 5, PCoIP image caching was introduced to minimize the bandwidth usage and there have been further enhancements in image caching in View 5.2. Hence, image caching improvements results are presented before discussing View 5.2 bandwidth results.

PCoIP Image Caching Improvements

As discussed in the earlier section, "View 5.2 New Features and Performance Improvements," there have been a number of optimizations in PCoIP image caching in View 5.2 and some related results are presented below.

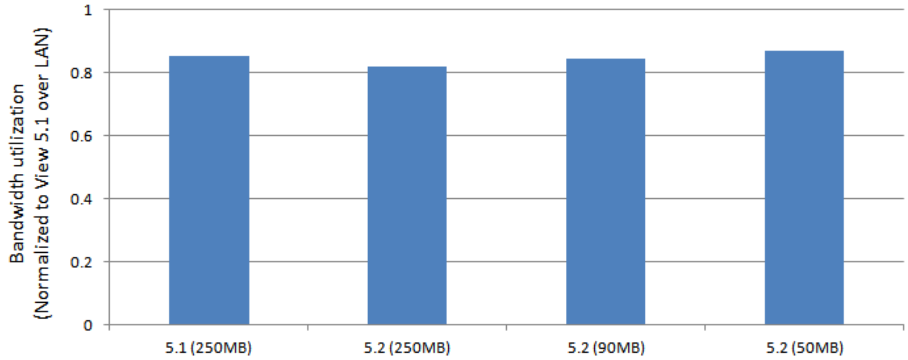

Figure 4. Normalized bandwidth utilization of View 5.1 and View 5.2 for different PCoIP image cache sizes

Figure 4 illustrates that, for typical office workflows, running View 5.2 with up to a 5X smaller cache can still deliver significant bandwidth savings; a 90MB View 5.2 cache was found to deliver comparable performance to View 5.1 configured with a 250MB cache, and even a 50MB View 5.2 cache delivered the majority of the bandwidth reduction benefits observed from View 5.1 configured with a 250MB cache. This up to 5X reduction in cache size can be a compelling option for memory-constrained thin clients or tablet devices. The maximum image cache size can be configured using GPOs or set on the client device.

Alternatively, users can continue to leverage the default 250MB cache size in View 5.2 and will see reduced bandwidth utilization in comparison with View 5.1, which is discussed in the next section.

Bandwidth Improvements in View 5.2

PCoIP in View 5.2 supports better image caching, which results in more image cache to store the display data so that bandwidth usage can be further minimized. To illustrate this, Figure 5 presents the bandwidth usage of View 5.1 and View 5.2 for different network conditions.

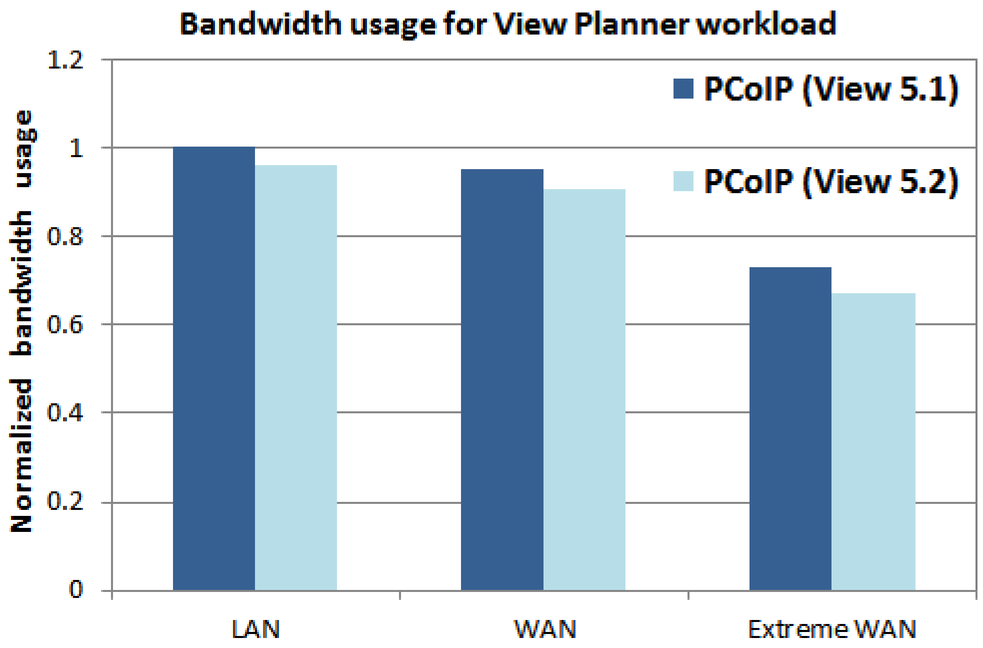

Figure 5. Normalized bandwidth usage for View 5.1 and View 5.2 for View Planner workload (10 seconds think time) for different network conditions (LAN, WAN, extreme WAN).

The y-axis shows the normalized bandwidth usage (lower is better) where the normalization is done with respect to View 5.1 bandwidth usage in a LAN network. Results show a 5-10% bandwidth reduction in View 5.2 for a typical office user workload for different network conditions. These reductions can be compelling when consolidating View sessions from a branch office onto a limited capacity link, or when users are connecting over congested WiFi connections. Furthermore, as would be expected, reducing the number of image blocks being encoded not only reduces the bandwidth utilization, but also has the benefit of improving interactivity (faster transmission of updates and the opportunity for higher frame rates, given the reduced bandwidth utilization) and reducing CPU consumption (less encoding work being done).

Bandwidth Usage for Windows 8 Desktop

This section presents the PCoIP bandwidth for a Windows 8 desktop VM. Due to new user experience enhancements, out-of-the-box Windows 8 has several animations which can significantly add to the overall bandwidth usage. These animations can be suppressed automatically using the group policy settings or manually in the parent desktop template.

One such setting is to disable the animations during Windows Maximize or Minimize. In Windows 8, Windows Maximize and Minimize operations cause a lot of intermediate display updates for better visualization; however, in a remote environment with bandwidth constraints, this can feel quite sluggish. So, it is strongly recommended to turn off the animation using the visual effects setting shown in Figure 6 on the left side.

Another finding in Windows 8, specifically with new version of Internet Explorer (IE 10), is that the cursor might remain in the waiting or busy state even after the page is loaded. These cursors are usually animated and just to transfer the cursor updates may consume up to 300 kbps of bandwidth. Hence, it is recommended to use the default cursor in bandwidth-constrained environments. Figure 6 shows the optimal settings.

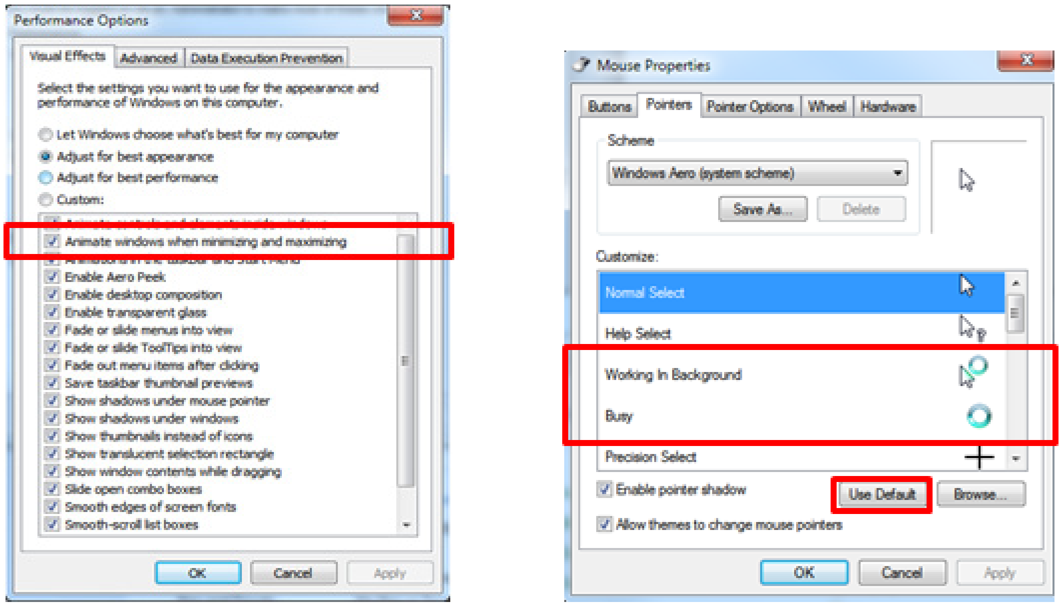

Figure 6. Windows 8 Optimizations (a) Disable animations for windows maximize/minimize operation (shown on the left side) (b) Use default cursor for busy and working cursors (shown on the right side)

Next, how these optimizations affect the overall bandwidth usage in Windows 8 is presented. Figure 7 shows the normalized bandwidth usage of a Windows 8 desktop for the View Planner workload with PCoIP for typical LAN network conditions. As seen, disabling animation effects during maximize and minimize operations can provide a 40% savings in the bandwidth usage. Also, changing the animated cursor to use the default cursor can provide an additional 20% bandwidth savings.

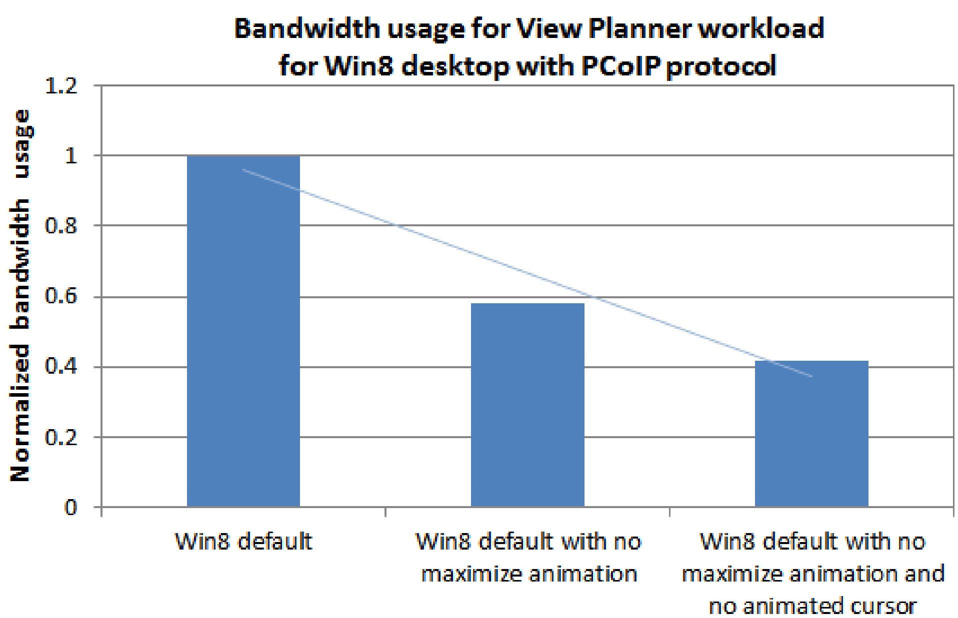

Figure 7. Normalized bandwidth usage for Windows 8 comparing different optimizations with out-of-the-box settings in a LAN network condition

RDP 8 Performance

RDP 8 is available by default in Windows 8 and can now also be achieved in a Windows 7 desktop by applying some patches. In this section, RDP 8 performance is compared with PCoIP when used with a Windows 7 desktop and a Windows 8 desktop.

Figure 8 shows the normalized bandwidth usage for the View Planner workload with PCoIP and RDP8 protocols with a Windows 7 and a Windows 8 desktop VM in a LAN environment. With the Windows 8 optimizations discussed in the previous section, Windows 8 bandwidth usage for both PCoIP and RDP8 is clearly lower compared to Windows 7. When RDP 8 is compared with PCoIP in Windows 7, there is 10% to 15% lower bandwidth usage with the PCoIP protocol. When the same comparison is done in Windows 8, about 15% to 20% lower bandwidth usage is seen with PCoIP.

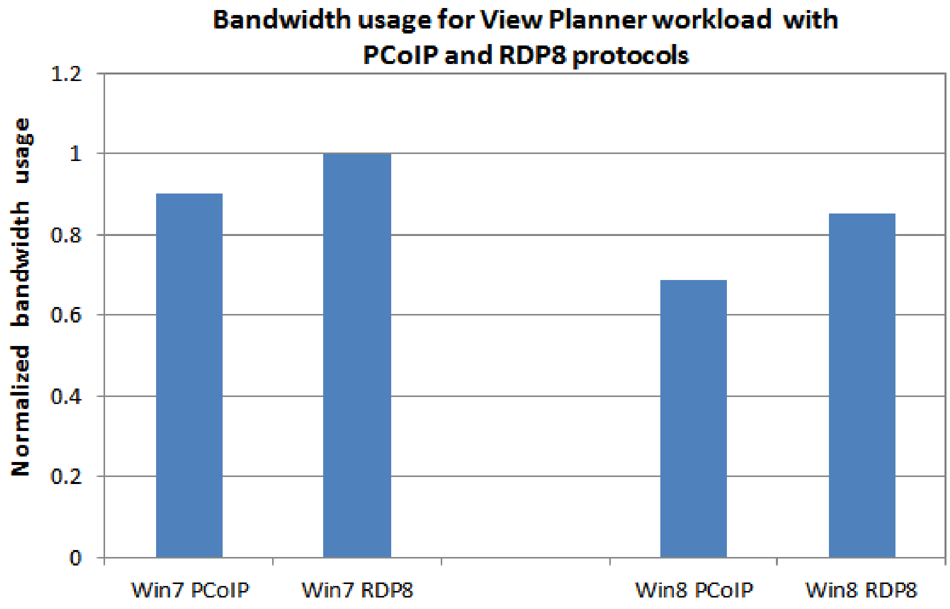

Figure 8. Normalized bandwidth usage for PCoIP and RDP 8 for LAN environment

Host Consolidation

The experimental setup shown in Figure 9 was used for the host consolidation experiments. A Dell PowerEdge R710 blade server with a 2.2GHz Intel Xeon E5-2660 processor and 392GB physical memory hosted the desktop VMs. The desktop VM's data and OS disk resided on a local SSD disks and the VM ran Windows 7 32-bit and Windows 8 32-bit with 1 vCPU and 1GB of virtual memory. The desktop screen resolution was set to 1152 x 864 with true-color mode. To run clients at scale, VMs were used on a separate host. A Dell PowerEdge R710 blade server with a 2.4GHz Intel Xeon E5645 processor and 296GB physical memory hosted the client VMs. Windows XP 32-bit clients were used to connect to the desktop VMs with the PCoIP protocol and the clients were configured with 1 vCPU and 768MB of virtual memory.

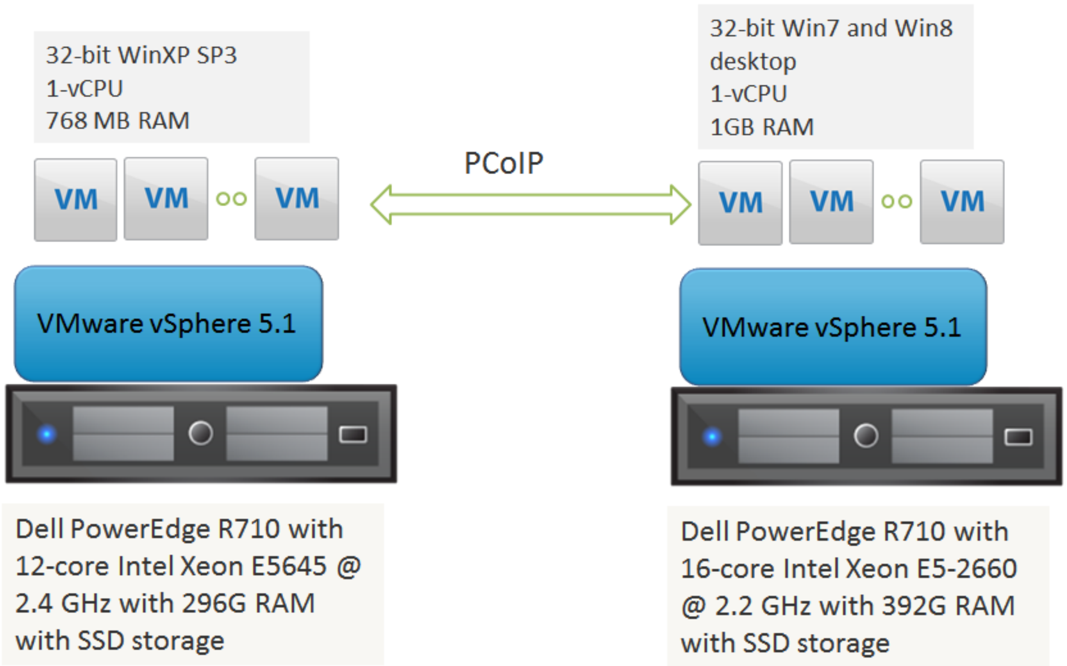

Figure 9. Experimental setup for finding the host consolidation with View Planner workload

The VMware View Planner workload was run on all the desktop VMs where all the applications were selected and a think time of 10 seconds was used. A 10-second think time represents operations happening almost at twice the rate than with a 20-second think time, so consolidation will be slightly less in this case. The View Planner workload with a 10-second think time was run on 192 Windows 7 desktops VMs on the 16-core Dell blade and the 95% View Planner response time was within the acceptable threshold of 1.5 seconds. Hence, 12 Windows 7 VMs per core can be loaded on this system for this workload set and the same results were also seen for Windows 8 VMs.

Hardware Accelerated 3D Graphics (vSGA) Performance

The aim of this vSGA performance whitepaper is to demonstrate the scalability of a VDI solution that uses vSGA to support 3D graphics. Accordingly, the paper focuses on four different workloads that stress the vSGA solution in different ways. VMware View Planner was used to measure the scalability—defined in terms of the consolidation ratio and the corresponding response time or frame rate during the runs. The chosen workloads represent typical customer use scenarios. The two workloads are:

- **Light 3D Workload:** This workload is composed of common desktop applications, including Office 2010, Adobe Acrobat, 720p video, IE9x static content, IE9x displaying a Web album, and Google Earth running in the Chrome browser. All these applications are launched at the beginning of the run and remain open for the duration of the run. Throughout the duration of the test, the workload performs a variety of different operations using these applications. The exact ordering of the operations differs from desktop to desktop to mimic real-world workloads. The desktop VMs run Windows 7 at a resolution of 1600x1200 pixels and have Aero enabled. This workload represents a use-case scenario typical of a knowledge worker.
- **CAD Workload 2:** A Solid Edge CAD viewer is run in isolation for the duration of the test. During the test, a 3-1 reducer model was used.

In initial performance testing, it was quickly discovered that the sophisticated image caching techniques in View 5.2 ensured that any repetitive interaction with the CAD applications was rapidly cached such that, in some cases for the remainder of the test, View was able to source up to 90% of the total remoted pixels from the image cache. Accordingly, simple model rotations or model animations are not suitable operations for examining the real-world performance of the system. Time was spent devising a more real-world interaction with the 3D models. The goal was not to completely defeat the View image caching, but to manipulate the model in a way that more closely mimics the potential usage by an actual CAD user. After studying how users tend to interact with 3D models, an automated interaction with the model was devised that approximates this process and this method is used in the CAD workloads that are presented in this whitepaper.

The tests use a modified version of View Planner 2.1 that supports higher 1600x1200 screen resolutions and incorporates 720p video streams. A "think time" of ten seconds was chosen for the View Planner runs, and the CAD workloads were implemented using View Planner's support for custom apps.

Finally, all of the performance tests discussed in this white paper were run with mostly default settings, except for ensuring that the registry setting "MaxAppFrameRate," described in the section titled "Optimizing Resource Sharing", was set to 30 frames per second to match the default PCoIP frame rate.

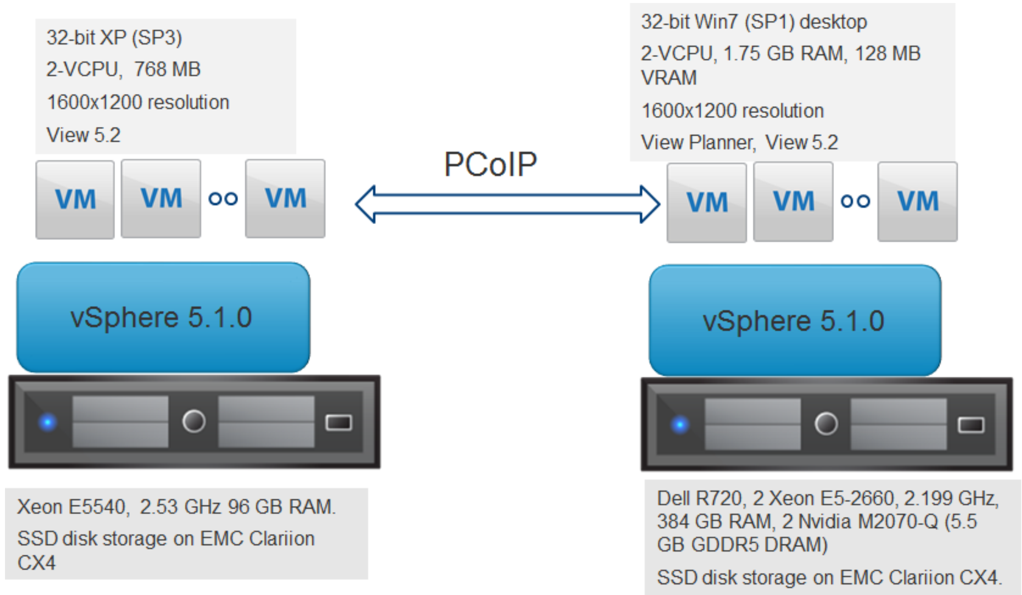

Figure 10. Experimental setup for measuring the performance of the vSGA stack and View 5.2.

The scalability of VMware's VDI solution was investigated by running these workloads on a single Dell R720 server with different VM consolidation ratios. The number of VMs that can be supported per GPU can be dictated by either the GPU's compute resources being exhausted or the GPU's available memory being exhausted. For the light 3D workload, it was found that GPU memory is the limiting factor on the GPU. Accordingly, the test system leveraged 2 GPUs, each with almost 6GB of RAM, to maximize the number of VMs that could be supported. When running with 2 GPUs, the major bottleneck to further scaling was found to be the CPU. CAD workloads were experimented upon with both 1 GPU and 2 GPU setups. This testing found that, in both scenarios for the dual-socket server under test, the CPU represented the bottleneck to further scaling.

The CAD workloads and light CAD workloads stress the CPU and show how well the vSGA solution scales with CAD workloads running at peak load either in isolation or together with other applications.

Light 3D Workload

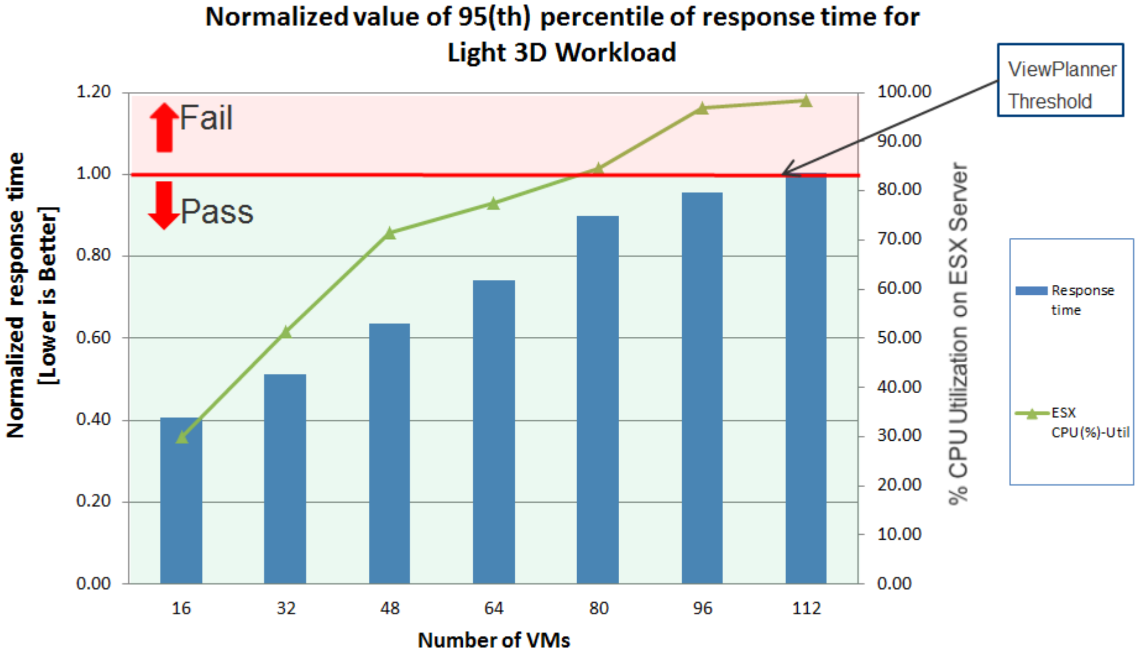

Figure 11. The bar chart presents the normalized values of the 95th percentile of the response time as the number of VMs is increased. Scaling is continued until performance falls below the View Planner response time requirements. The corresponding CPU utilization, as measured using `esxtop`, is shown by the line graph. The peak GPU utilization was 20%.

In this initial test the light 3D workload was run using View Planner and the number of VMs gradually increased until the View Planner response threshold was exceeded. These results are presented in Figure 11. Based on this data, it is clear that the vSGA stack can support 7 users per core on this system while each user is executing the light 3D workload; running on higher performance processors will typically deliver even higher consolidation ratios.

As discussed previously, the results shown in Figure 11 were obtained using desktop VMs configured with 128MB VRAM. Since the test bed had two Nvidia M2070-Q GPUs with about 5.5GB DRAM each, only around 80 desktop VMs can be supported by the GPUs. As the number of VMs is increased further, the additional VMs are supported using the software-renderer. For this light 3D workload, the maximum consolidation ratio achieved on the dual-socket server under test was 112: the test was stopped when, at 112 VMs, the VM responsiveness reached the upper limit allowed by the View Planner responsiveness threshold. At this point, both the GPU memory and CPU resources of the system are exhausted and further consolidation, while maintaining acceptable user experience, is not possible.

User Experience

To both accurately and automatically quantify and measure user experience at scale, VMware has developed a test suite composed of six tests: three simple 3D rendering tests, a fourth test that drags a window across the screen, a fifth that scrolls the contents of a window, and a final test that maximizes and minimizes a window, all of which are controlled by mouse events sent by the View client. These tests are constructed using patent-pending techniques that enable performance agents installed on the View clients to accurately identify each of the frames that are rendered to the View client's display. The performance agent installed on the View client leverages the frame identity information, in conjunction with its knowledge of when each frame was actually rendered to the clients to compute a user experience metric. This metric can be used to quantitatively characterize the remote

VMs responsiveness in the response to mouse events as the consolidation ratio is increased.

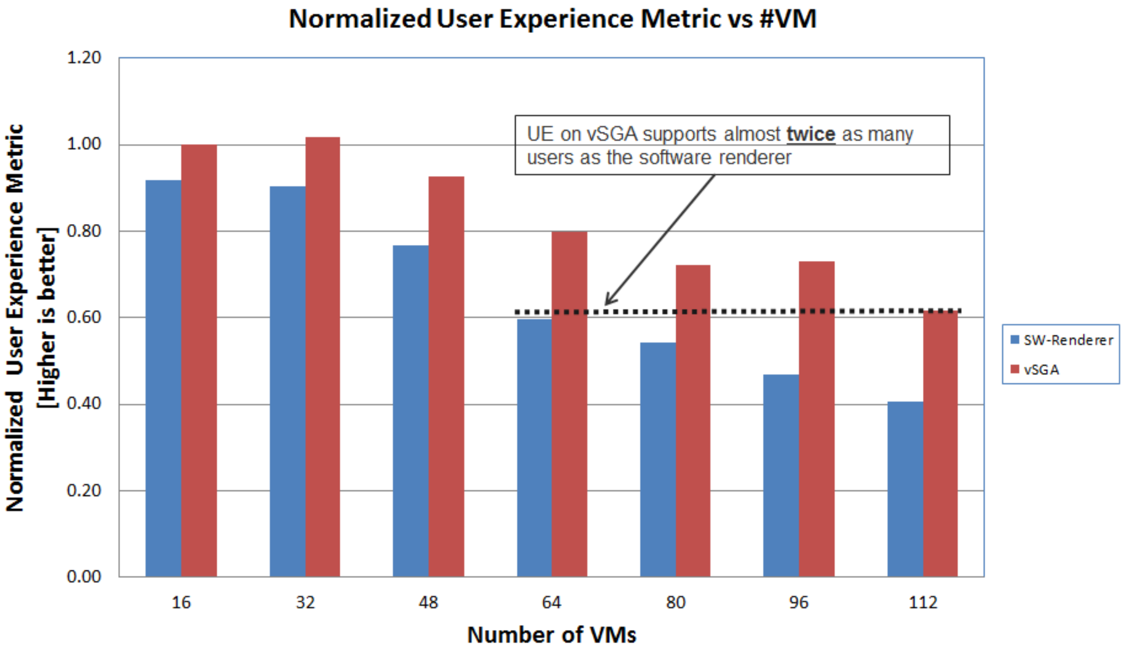

Figure 12. Presents normalized user experience metrics for both the vSGA and software renderer solutions as the number of desktop VMs is increased. The data clearly illustrates that the user experience observed with 112 users on a vSGA based system is equal to that observed with only 64 users when using a pure software renderer solution.

User experience was measured while running the light 3D workload on both vSGA and the software renderer. The normalized values of user experience are shown in Figure 12, and clearly illustrate:

- Using hardware accelerated 3D improves responsiveness in comparison with a software solution, even at low consolidation ratios, where CPU is not exhausted.
- Adding GPUs to an existing software-renderer solution enables the VM consolidation ratio to be almost doubled while maintaining user experience.

CAD Workload (Solid Edge Viewer)

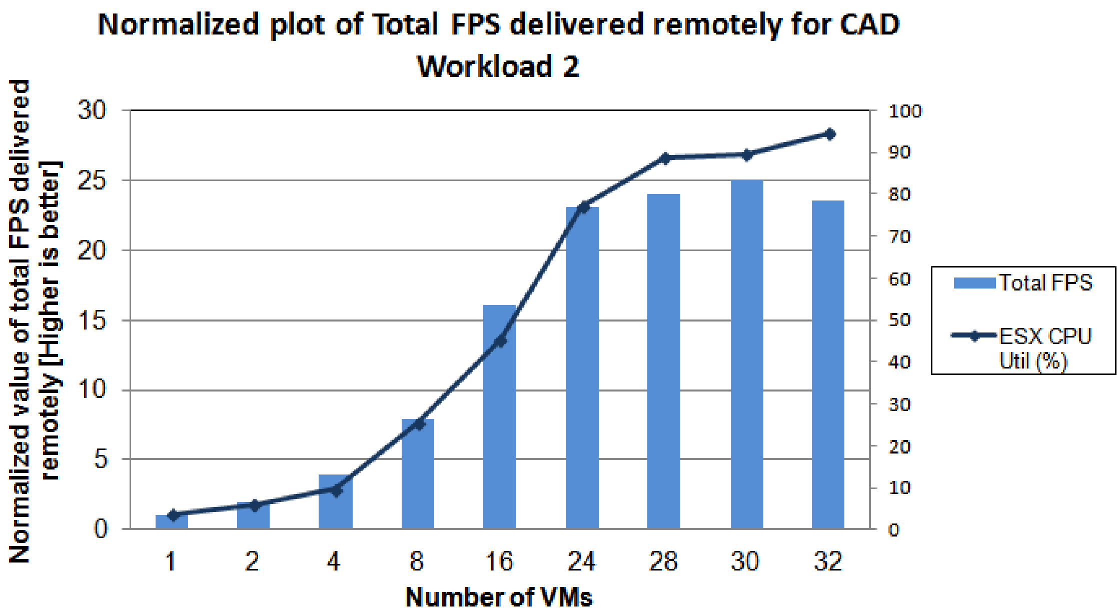

Figure 13. This bar chart presents the scalability of the vSGA solution as the load on the server is increased. The results are normalized and the frame-rate observed with just one VM running on the server is defined as the basis for comparison. The corresponding CPU utilization as measured using `esxtop` is shown by the line graph. GPU utilization was observed to be 85%.

In this workload, the Solid Edge viewer replaces the SolidWorks viewer and runs a single model: a 3-to-1 reducer. As with the previous CAD workload, the simulated user's interaction with the model is designed to mimic a real user's potential usage pattern. Figure 13 illustrates the scalability of the vSGA and View solution; showing the aggregate remoted frame-rates (FPS) as the number of desktop VMs on the server is steadily increased. The results are fairly similar to that observed with CAD Workload 1; as the number of VMs is increased from 1 to 32 the aggregate remoted frame-rate increases by 25X. One difference to note between the results for CAD Workload 1 and CAD Workload 2 is the peak GPU load observed: With workload 1, it peaked around 37%, whereas with Workload 2, it reached as high as 85%. This difference results from the differing complexity of the 3D models, but highlights that the vSGA solution scales equally well when the GPU is close to fully utilized.

Figure 14 presents the same results as Figure 13, but presents the remoted frame-rate data on a per-VM basis. This view of the data highlights that the performance of the individual VMs sees little decrease as the number of VMs is scaled. This is further emphasized by the standard deviation results shown on the bar chart in 4: not only does the average VM performance remain pretty constant as the number of VMs is increased, but performance is shown to be very consistent across all the VMs. This clearly highlights the ability of ESXi to fairly share the resources between the VMs, even under load and ensure that the user experience is consistent across VMs.

. .

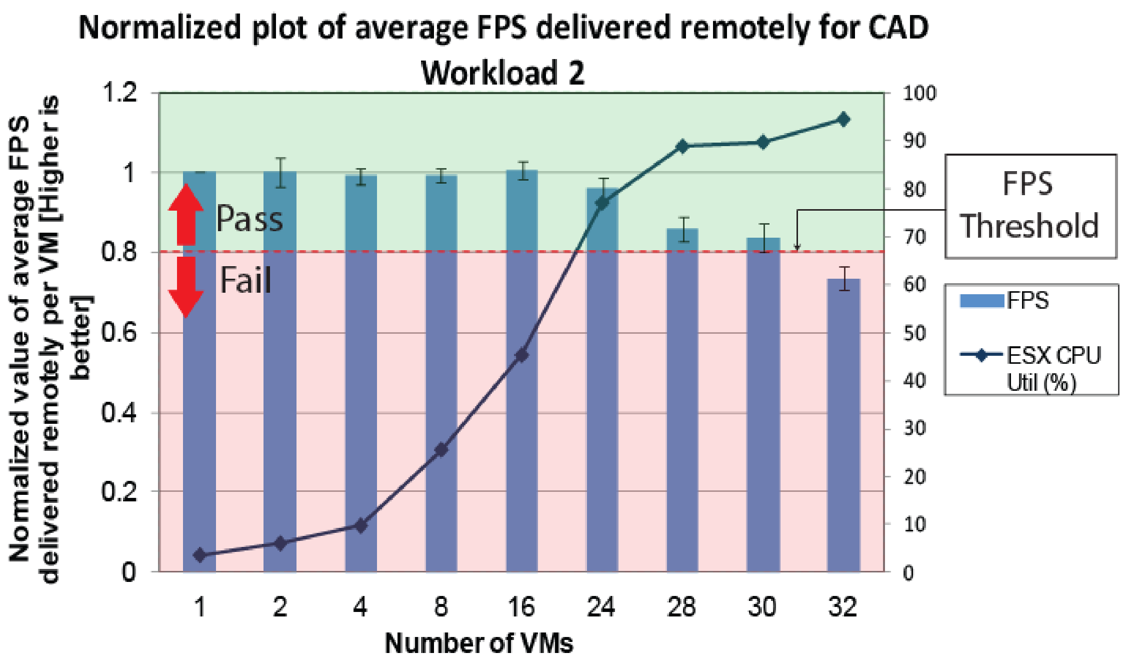

Figure 14. The bar chart presents the average per VM remoted frame-rates (FPS) observed with the Solid Edge viewer as the number of VMs is increased. The results are normalized, with the frame-rate observed with a single VM used as the basis for comparison. In addition, the bar chart is also marked with the associated standard deviation. The corresponding CPU utilization as measured using `esxtop` is shown by the line graph.

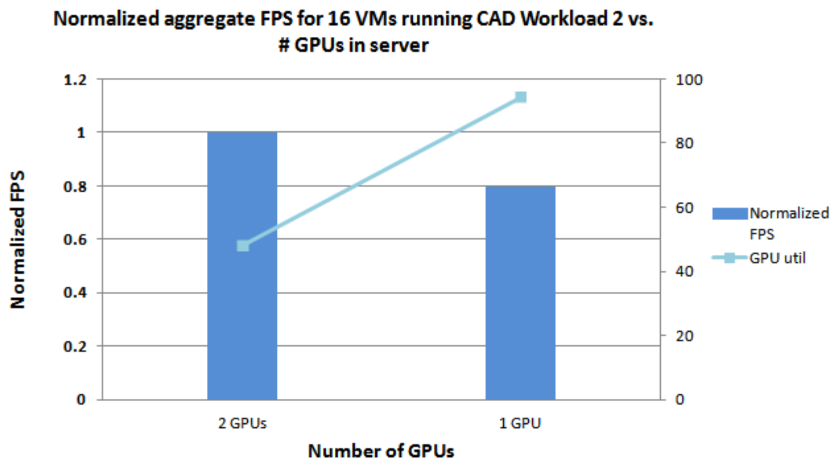

Figure 15. The bar chart presents the remoted frame-rates observed with 16 desktop VMs when running with a single GPU and then with 2 GPUs. The results are normalized with the frame-rate observed with a single VM used as the basis for comparison. The corresponding GPU utilization is shown using the line chart.

Finally, as discussed previously, the server in the test-bed used to host the desktop VMs was equipped with two GPUs in order to support the GPU memory requirements of the light 3D workloads. However, for the CAD workloads, the available GPU memory is not a bottleneck. Accordingly, to demonstrate the performance of vSGA with just one GPU, the CAD workload 2 was also run with 16 VMs and a single GPU. The results are shown in Figure 15 and clearly illustrate that, while the GPU utilization almost doubles with one GPU (as would be expected), the remoted frame-rate falls by only 20%, even though the single GPU is almost 100% utilized.

Space Efficient Sparse (SESparse) Disk Performance

SESparse can now reclaim spaces from the linked clone pools and provides efficient usage of available disk space. To evaluate the SE Sparse disk performance, some custom apps were implemented in the View Planner. These custom apps mimic the regular user behavior of installing and uninstalling applications, downloading and copying files, and then deleting some of these files. These custom apps along with some View Planner apps such as Office workload, Adobe Reader, and Web Album were ran for 2 iterations and a 10-second think time. Figure 16 shows the normalized disk space of 100 linked clone desktop VMs at different stages. After the View Planner workload was run, the space usage almost doubled when compared to their freshly provisioned state. In the View Planner workload, there were some uninstall and delete operations and hence, scope for reclaiming the space back. With the SESparse feature, by applying the wipe and shrink operation, most of the deleted and unmapped space can be reclaimed. In this experiment, the space reclamation was close to about 1GB per VM.

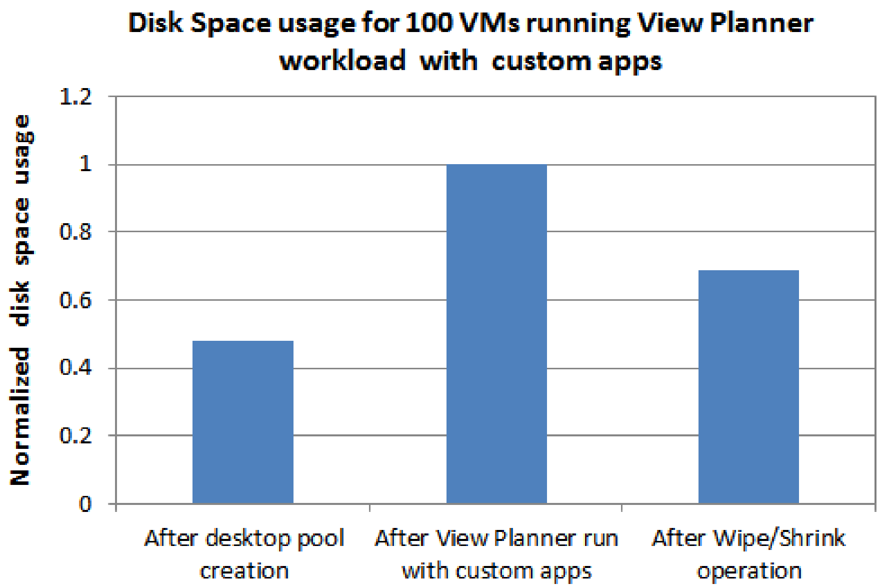

Figure 16. This figure shows the normalized disk space use of 100 VMs at different time periods: (1) When the pool is created and desktops are powered on. (2) After View Planner workloads with custom apps are run. (3) After the wipe/shrink operation. As clearly seen, the Wipe/Shrink operation can reclaim almost all the unused space (close to 1GB per VM in this experiment).

Since the wipe/shrink operation can be I/O-intensive for space reclamation, View administrators are encouraged to use the blackout periods appropriately (available in the View admin UI) to minimize any perturbation in the user experience. Also, depending upon the underlying storage, administrators can tune the concurrency level in LDAP (under OU=Properties, OU=Virtual Center) and edit the `pae-SeSparseOperationsLimit` for the desired vCenter.

View 5.2 Best Practices

Guest Best Practices

Optimizing the master VM is very important before the linked clones or full clones are created. Many optimizations can be applied to the guest VM; these optimizations will save precious resources such as bandwidth and storage.

PARAMETER	CONFIGURATION
vCPU	1 for WinXP and Win7 and Win8, 2 for multimedia intensive apps
Memory	512-768 MB for WinXP 1GB for 32-bit Win7 and Win8 2GB for 64-bit Win7 and Win8 1.5-2GB for WinXP, Win7, and Win8 32-bit 3GB for Win7 and Win8 64-bit for memory-intensive apps
Network adapter	vmxnet3, flexible
Storage adapter	PVSCSI or LSI Logic SAS
VMware Tools	Latest installed*
Visual settings	"Adjust to best performance" † Disable Animations for Windows Maximize and Minimize operations Use default cursor for busy and working cursor
Disable services ‡	Windows Update, Super-fetch, Windows Index
Group policy settings	Disable Hibernation System restore disable Screensaver to None
Other settings	Turn off clear-type Disable fading effects Disable auto-play and external drive caching for quick release§ Disable last access timestamps

Table 1. **Guest best practices to save bandwidth and storage resources**

Table 1 provides the settings that were applied to collect some of the single-VM results in order to achieve the best bandwidth and keep the need for storage resources as low as possible. More guest optimizations are discussed in the *Windows 7 Optimization Guide for VMware View* [4].

* Also make sure balloon driver is functioning properly.
† This can provide an additional 10-20% bandwidth savings in WAN environments.
‡ This is especially important to curtail redo log growth with linked clones.
§ This is recommended so that these services don't try to pull USB drive information over the WAN.

View PCoIP Best Practices

After optimizing the master VM, you can apply some PCoIP protocol best practices to realize the best user experience as shown in Table 2.

Table 2 lists some of the GPO settings that you can apply to improve the user experience for WAN environments. The first two settings in the table can be made regardless of network conditions—these changes will save lot of bandwidth with almost no discernible difference in quality.

SETTING	DEFAULT	RECOMMENDATION	DESCRIPTION
Build to lossless	On	Turn Off	Enables the ability to enable or disable build to lossless
Session Audio BW limit	500Kbps	50 - 100Kbps	Reduces bandwidth usage of audio with usable quality
Maximum frame rate	30	Change to 10-15 based on network settings	In WAN conditions, this will be helpful for video playback and fast graphics operations
Maximum session bandwidth	-	Set per network conditions	Good for better bandwidth estimation
Client side cache size	250MB	Set per client-side memory available	This allows you to configure the client side image cache size.

Table 2. **PCoIP protocol GPO settings and best practices**

- **Turn off build to lossless:** This is quite important for perceptually lossless quality (which is hard to differentiate from fully lossless quality for an Office user workload). This saves 20-30% of the available bandwidth.
- **Lower Session Audio BW limit:** We also suggest limiting the session audio bandwidth to be in the range of 50-100kbps and it will still provide the usable audio quality with significant lower bandwidth usage.
- **Set maximum frame rate and maximum session bandwidth according to the network conditions:** The default value of maximum frame rate is 30 and this setting can be lowered to anywhere between 10-15 for typical WAN conditions. This will provide decent video playback performance.
- **Set maximum session bandwidth according to your link bandwidth:** This will help maximize the user's experience of sharing the same link and will also help PCoIP to better estimate bandwidth. There are more PCoIP protocol settings which are discussed in View network optimization guide [7].

Best Practices

View 5.2 and PCoIP dynamically adapt to the available CPU and bandwidth resources to present the optimal user experience. Even when tens of VMs are sharing a single physical GPU, vSphere ensures that the resource is fairly shared between the different VMs. As a result, very little out-of-the-box configuration is required to deliver peak performance:

- Configure VMs to use VMXNET3 NICs.
- Consider disabling PCoIP's build-to-lossless mode.
- Set the MaxAppFrameRate (see "Optimizing Resource Sharing").

Conclusion

This paper presented best practices and performance data for VMware Horizon View 5.2 and also highlighted some of the new features. With optimized PCoIP image caching, a bandwidth improvement of up to 10% is seen across different network conditions. The other PCoIP improvements included dynamic GPO settings and the support of relative mouse. Windows 8 optimizations are also presented which can provide up to 60% bandwidth savings. There were also performance improvements in the View management side where up to 2x improvement is seen for various administrative operations such as provisioning, recompose, and rebalancing. Furthermore, with new SESparse disk, administrator can minimize storage overhead by efficiently using and reclaiming storage space to minimize the storage capacity requirements for persistent desktops and decreases the need to continuously recompose and restore images.

For the vSGA feature, the results illustrated the ability of VMware's hardware-backed 3D support to scale efficiently and it was illustrated that for light 3D workloads, a 2-socket x86 server with 2 GPUs can support over 100 3D desktops. This clearly showed the benefits of GPU virtualization and the strength of VMware's 3D strategy where VMs over-and-above those that can be supported by the GPUs seamlessly use VMware's software renderer. Finally, best practices for optimizing the parent template, for optimizing PCoIP for different network conditions, and for the vSGA feature were presented.

References

[1] *VMware View Planner 2.1 User Guide*. VMware, Inc., 2013.
http://communities.vmware.com/community/vmtn/servicessoftware/view_planner

[2] B. Agrawal, R. Bidarkar, S. Satnur, T. M. Ismail, L. Spracklen, U. Kurkure, V. Makhija, "VMware View Planner: Measuring True Virtual Desktops Experience at Scale." VMware Technical Journal, December 2012.
http://labs.vmware.com/academic/publications/view-vmtj-winter2012

[3] L. Spracklen, B. Agrawal, R. Bidarkar, H. Sivaraman, "Comprehensive User Experience Monitoring." VMware Technical Journal, March 2012.
http://labs.vmware.com/academic/publications/spracklen-vmtj-spring2012

[4] *VMware View Optimization Guide for Windows 7*. VMware, Inc., 2012.
http://www.vmware.com/resources/techresources/10157

[5] Remote Desktop Protocol 8.0 update for Windows 7 SP1. Microsoft, Last Review: Feb. 12, 2013, Revision: 8.0.
http://support.microsoft.com/kb/2592687

[6] "VMware Horizon View 5.2 and Hardware Accelerated 3D Graphics: Performance and Best Practices"
http://www.vmware.com/resources/techresources/cat/91,156,86

[7] *VMware View 5 with PCoIP: Network Optimization Guide*, VMware, Inc.
http://www.vmware.com/files/pdf/view/VMware-View-5-PCoIP-Network-Optimization-Guide.pdf

About the Authors

Dr. Banit Agrawal is a Sr. Performance Engineer at VMware. He has expertise and has filed several patents in the area of VMware View, remote display protocols, VMware View Planner, and performance troubleshooting.

Dr. Lawrence Spracklen is a Senior Staff engineer at VMware. His current research focus is VDI deployments, specifically protocol optimization and multimedia quality analysis.

Hari Sivaraman is a Sr. Member of the Technical Staff at VMware. He works on 3D rendering performance and on CUDA support on ESX.

Dr. Xing Fu is a Member of Technical Staff in the Performance Engineering group at VMware. His work focuses on the performance of virtualization solutions.

Rishi Bidarkar is Sr. Manager in the performance team at VMware. He leads the VDI Performance and View Planner teams. He has filed several patents in the area of VDI performance and display benchmarking.

Acknowledgements

Authors would like to extend their sincere thanks to Julie Brodeur for her comments, feedback, and improvements in the quality of this whitepaper. We also thank Jad Chamcham, Warren Ponder, Pat Lee, Rob Baesman, and Oswald Chen for their comments on the early draft.