VMware Horizon 6 and Hardware Accelerated 3D Graphics

Performance and Best Practices

TECHNICAL WHITE PAPER

Table of Contents

Introduction

The significant cost savings that can be realized by leveraging a virtual desktop infrastructure (VDI), coupled with the rapid growth in cheaply available bandwidth (both on LAN and WAN), means that VDI no longer needs to be constrained by low-resolution desktops associated with task workers, but can be leveraged to cost effectively bring high-fidelity, high resolution, and multi-monitor 3D desktops to a wider audience.

This whitepaper discusses the support for hardware accelerated 3D graphics that is available with VMware vSphere® 5.5 and VMware Horizon™ 6 and presents performance and consolidation results for a number of different workloads, ranging from simple3D desktop usage to performance-intensive CAD-based workloads. Further, given that the intensity of a 3D workload will vary greatly from user to user and application to application, rather than highlighting specific case studies, this paper demonstrates how the solution efficiently scales for both light- and heavy-weight 3D workloads. This paper also presents key best practices to extract peak performance from a 3D VMware Horizon 6.0 deployment.

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 Horizon 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 Horizon 6.0 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 Horizon 6.0 to include key technologies such as CAD and medical imaging.

3D Desktops in Horizon 6

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. For complete details on managing and configuring 3D desktops, refer to "VMware Horizon 6 Documents" [1].

Supported Configurations

Currently, vSphere supports DirectX 9 and OpenGL 2.1 applications running in both Windows 7 and Windows 8 VMs. vSphere 5.5 provides virtualization support for a range of Nvidia GPUs. Currently, the supported GPUs are Grid K1 and Grid K2.

Best Practices

The PCoIP protocol 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.
- In Horizon 6, PCoIP build-to-lossless mode is disabled by default, so you don't need to set this explicitly.

The next sections describe additional considerations for maximizing performance and efficiency of a 3D desktop workload.

Maximizing VM Consolidation

The administrator console provides the ability to configure the amount of graphics or video RAM (termed VRAM) allocated to each VM. The default per-VM VRAM allocation is 128MB. While increasing the per-VM VRAM allocation might deliver higher performance, it will limit the number of VMs that can simultaneously share a GPU; the GPU's memory is sub-divided between the VMs. Accordingly, if the GPU has 4GB of memory, and the VMs are configured to use the default 128MB of VRAM, it is possible for up to 32 VMs to simultaneously share the GPU. Similarly, if the VMs are configured to use 512MB each, the per-GPU consolidation ratio may be reduced to 8. To further increase the number of VMs supported per server, vSphere 5.5 can support multiple GPUs per system.

For Aero and basic 3D applications, use of the default VRAM allocation is sufficient and is recommended to ensure maximum consolidation ratios. Higher allocations can be reserved for VMs that will run more intensive 3D operations, where additional GPU resources will deliver improved performance.

For pools containing more VMs than can be simultaneously supported by the GPU, additional VMs will not boot once the GPU resources have been exhausted. In this situation, rather than explicitly setting the pool to use hardware 3D in the administrator console, it is advantageous to use the "automatic" option; this enables any additional VMs that might be needed, over and above the GPU's capabilities, to be supported using vSphere's software renderer solution.

In contrast, for situations where two distinct groups of users share a server—one group requiring hardware 3D acceleration and one not—it is best to configure these two groups of VMs as separate pools. Use the administrator console to explicitly configure the hardware group to use hardware accelerated 3D and the other group to use either software 3D or even no 3D, as appropriate. Changes to a pool's 3D strategy are handled automatically by Horizon View and vSphere and do not require manual configuration of the desktop VMs.

Optimizing Resource Sharing

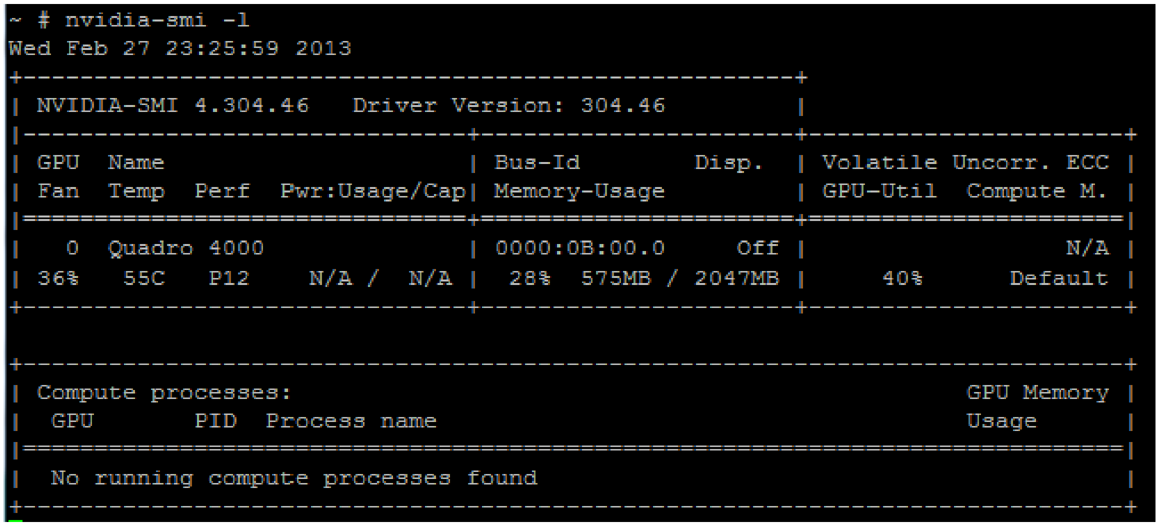
In contrast to a physical workstation that has sole use of its GPU, in the virtualized environment GPUs become a shared resource. As a result, it is important to ensure that each VM does not use the GPU resource in a wasteful manner. For instance, in many situations it often does not make sense for a 3D application to render hundreds of frames per second if Horizon is configured to remote at a lower frame-rate (30fps is the default setting). For these situations, Horizon 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.

Tracking Compute Resources

When consolidating multiple GPU accelerated 3D VMs onto a server, it is important to track both CPU and GPU utilization. When considering a VM's CPU utilization, it is important to consult host-level information in order to ensure that the contributions of all virtual machine components are considered. This is readily achieved either by using `esxtop` or by consulting the appropriate graphs in the vSphere client or vCenter. GPU utilization can be determined by leveraging `nvidia-smi` on ESXi. This command returns information, as illustrated in Figure 1, showing the GPU's memory usage and utilization.



```
~ # nvidia-smi -l
Wed Feb 27 23:25:59 2013
+------------------------------------------------------+
| NVIDIA-SMI 4.304.46   Driver Version: 304.46         |
|-------------------------------+----------------------+----------------------+
| GPU  Name                     | Bus-Id        Disp.  | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap| Memory-Usage         | GPU-Util  Compute M. |
|===============================+======================+======================|
|   0  Quadro 4000              | 0000:0B:00.0     Off |                  N/A |
| 36%   55C   P12    N/A /  N/A |  28%  575MB / 2047MB |     40%      Default |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Compute processes:                                               GPU Memory |
|  GPU       PID  Process name                                     Usage      |
|=============================================================================|
|  No running compute processes found                                         |
+-----------------------------------------------------------------------------+
```

Figure 1. Using `nvidia-smi` to check GPU utilization in ESXi

By examining both the GPU and CPU utilization, it is simple to determine what, if any, resource is first exhausted, and begins to restrict further scaling.

Client Sizing

3D workloads are frequently more graphically intensive than traditional office applications. As a result, they have the capacity to place higher computational demands on the user's client device. Accordingly, some thin clients may lack the computational resources to deliver a high-quality 3D experience to the end user. While the exact client device requirements are driven by the specifics of the workload running on the remote desktop, for the workloads undertaken in this white paper, a single-core Intel Atom-based thin client was found to be adequate with significant idle CPU time still available.

vSGA Performance Tests

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 [2] (version 3.5) 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 four 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 office worker.
- **Light CAD Workload:** This workload adds the SOLIDWORKS CAD viewer to the "Light 3D Workload." The SOLIDWORKS CAD viewer is typical of applications used by a CAD content consumer, and this workload represents the use case where CAD viewers are used occasionally in conjunction with typical office applications. In this test, the SOLIDWORKS CAD viewer is used to run two models: a sea-scooter (as illustrated Figure 2, upper left) and a cross-section of a shaft (as illustrated in Figure 2, **upper** right). The models are run sequentially using a single SOLIDWORKS viewer.

- **CAD Workload 1:** The SOLIDWORKS CAD viewer is run in isolation; that is, without any other applications for the duration of the test. This workload uses the same sea scooter and cross-section models to demonstrate the capability of the system to generate and remote frames when using these models in a manner that is typical of the use-case scenario for a CAD content consumer.
- **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, as illustrated in Figure 2, bottom. This workload is run in a manner similar to that of "CAD Workload 1."

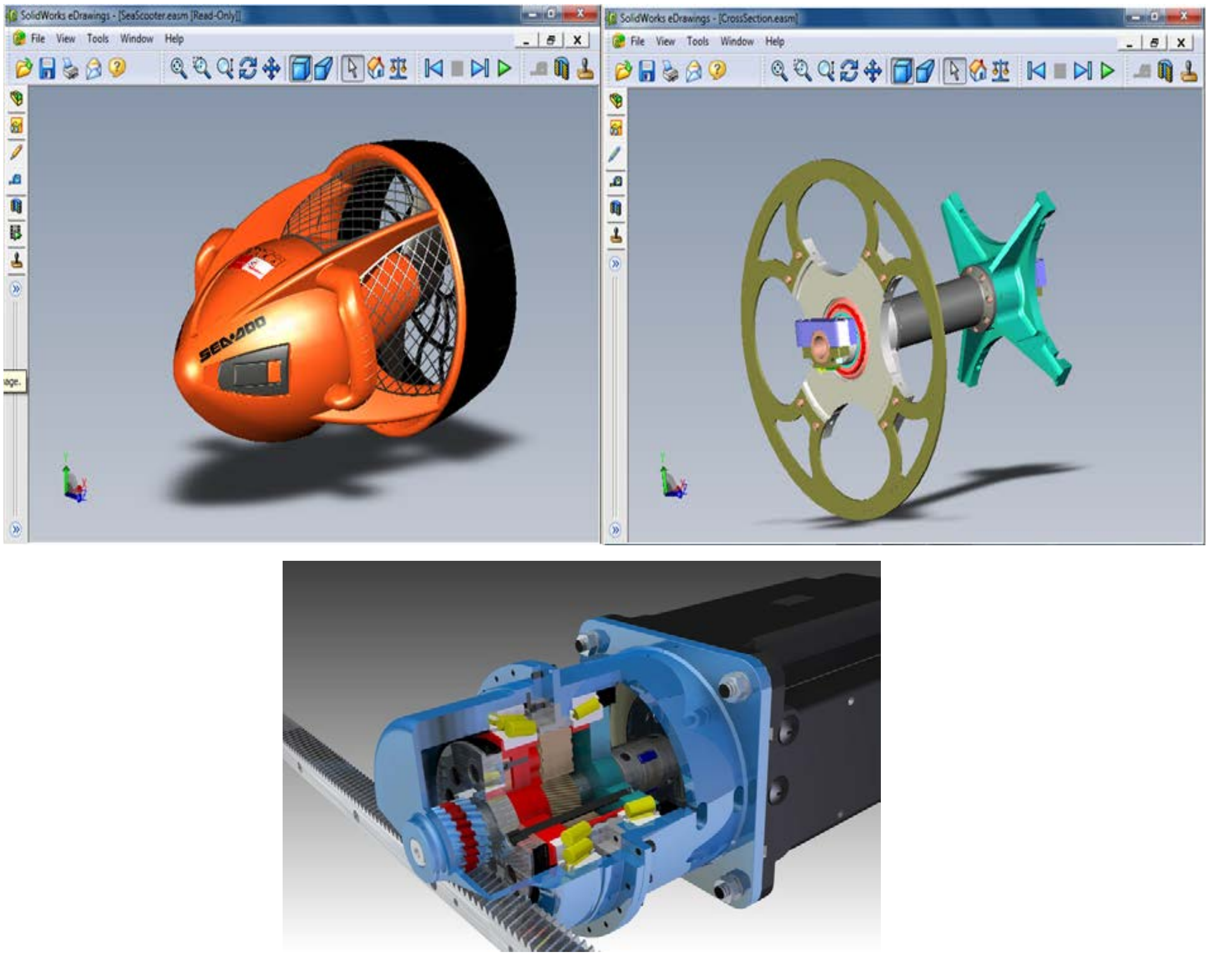

SolidWorks eDrawings - [SeaScooter.easm [Read-Only]]
File View Tools Window Help
SolidWorks eDrawings - [CrossSection.easm]
File View Tools Window Help
SEA-DOO

Figure 2. Shows the "SeaScooter" (top left) and "CrossSection" (top right) models that were used in performance tests with the SOLIDWORKS CAD viewer and the "3-to-1 reducer" (bottom) model that was used in performance tests with the Solid Edge Viewer.

In initial performance testing, it was quickly discovered that the sophisticated image caching techniques in Horizon 6.0 ensured that any repetitive interaction with the CAD applications was rapidly cached such that, in some cases for the remainder of the test, Horizon 6.0 was able to source up to 90% of the total remotely delivered 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 Horizon 6.0 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 white paper.

Test Bed Architecture

The test bed architecture, shown in Figure 3, uses the View Planner architecture [2], and is composed of three major logical components:

- The workload is executed on Windows 7 VMs. These VMs are referred to as the *desktop VMs* and each one is configured with 2 vCPUs, 1.75GB of memory, and 128MB of Video RAM. The desktop VMs are all located on a single physical server as shown in Figure 3. The load on the server and the GPUs is varied by changing the number of desktop VMs.
- The simulated users use VMware Horizon 6.0 software clients running on Windows 7 VMs to connect to their Windows 7 desktop VMs. These Windows 7 VMs are referred to as *client VMs*, and, as illustrated in Figure 3, are located on a second physical server. The simulated users use 2-vCPU VMs as their client machines. Guidance on client sizing is discussed in the "Best Practices" section.
- The AD server, the View Planner appliance, the VMware View-broker, and VMware vCenter are run using VMs located on a third physical server, not shown in Figure 3.

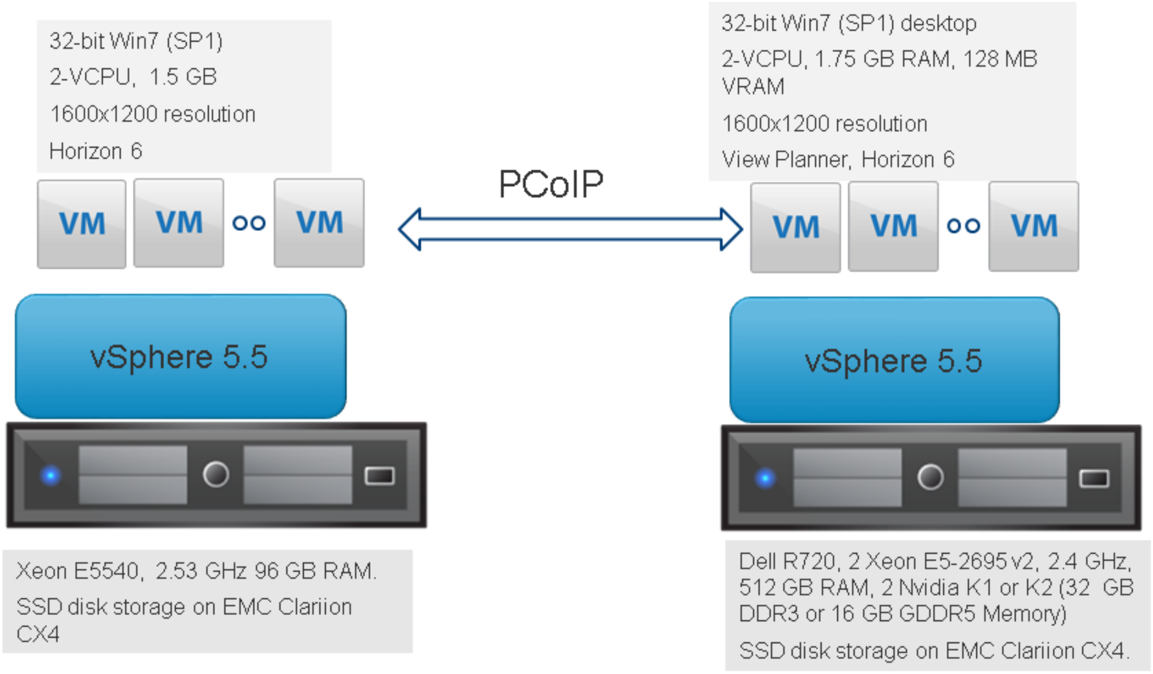

32-bit Win7 (SP1)
2-VCPU, 1.5 GB
1600x1200 resolution
Horizon 6
32-bit Win7 (SP1) desktop
2-VCPU, 1.75 GB RAM, 128 MB VRAM
1600x1200 resolution
View Planner, Horizon 6
PCoIP
VM
VM
oo
VM
VM
VM
oo
VM
vSphere 5.5
vSphere 5.5
Xeon E5540, 2.53 GHz 96 GB RAM.
SSD disk storage on EMC Clariion CX4
Dell R720, 2 Xeon E5-2695 v2, 2.4 GHz, 512 GB RAM, 2 Nvidia K1 or K2 (32 GB DDR3 or 16 GB GDDR5 Memory)
SSD disk storage on EMC Clariion CX4.

Figure 3. Experimental setup for measuring the performance of the vSGA stack and Horizon 6.0.

The tests use VMware View Planner 3.5. A “think time” of five seconds was chosen for the View Planner runs, and the CAD workloads were implemented using View Planner’s support for custom apps. All of the performance tests discussed in this white paper were run with default settings.

Performance Metrics

For VDI deployments, performance is typically measured by the number of users that can be supported with a certain level of remote desktop *responsiveness*. Responsiveness is defined by a variety of metrics, including application response time and remotely delivered frame-rates. The following metrics are used to quantify Horizon 6.0 performance:

1. **Consolidation ratio:** The number of users that can be supported concurrently on a server. The consolidation ratio is also frequently presented as the number of VMs per processor core.
2. **Response time:** View Planner measures the response time for non-I/O-bound operations while the workload is running and reports the 95th percentile of these response times. In this white paper, the results are normalized to the maximum allowable View Planner threshold [2] response time.
3. **Remotely delivered frame rates:** VMware has patent-pending techniques that accurately measure the number of remotely delivered frames that correspond to frame updates generated by 3D applications and videos.

We report the consolidation ratios achieved in conjunction with either response time or remotely delivered frame-rate metrics. As a result, there is no single consolidation ratio, rather a range of consolidation ratios that can be reconciled with a user’s perception of acceptable performance. For the typical office workloads run by View Planner, based on extensive testing, a View Planner threshold [2] response time was found that represents the upper limit on an acceptable response time. A detailed description of this selection process is available in the View Planner discussion forum [3]. In this paper, the maximum consolidation ratio that can be achieved is represented by the maximum number of VMs that can be run without violating the View Planner threshold.

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 and the light CAD workload, the View Planner threshold response time was used as the stopping criteria.

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. In the case of the CAD workloads, the number of frames that could be remotely delivered was the limiting factor; in most cases, once that fell below a threshold value, the number of VMs were stopped from scaling up.

Scalability Results

Light 3D Workload

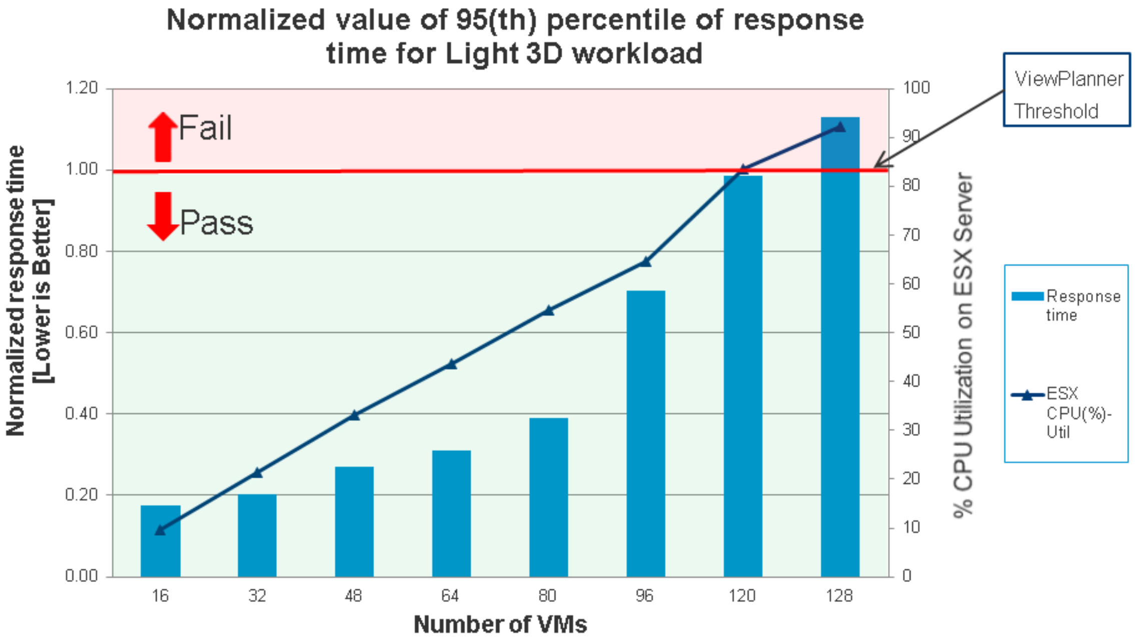


Figure 4. The bar chart presents the normalized values of the 95th percentile of the response time as the number of VMs is increased. Scaling is continued until performance falls below the View Planner response time requirements. The corresponding CPU utilization, as measured using esxtop, is shown by the line graph.

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 4. Based on this data, it is clear that the vSGA stack can support 120 users 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 4 were obtained using desktop VMs configured with 128MB VRAM. Since the test bed has two Nvidia K2 GPUs with about 8GB DRAM each, only around 128 desktop VMs can be supported by the GPUs. For this light 3D workload, the maximum consolidation ratio achieved on the dual-socket server under test was 128: the test was stopped when, at 128 VMs, the VM responsiveness exceeded the upper limit allowed by the View Planner responsiveness threshold.

Light CAD Workload

As discussed in the initial workload descriptions, the light CAD workload is run using View Planner and involves adding SOLIDWORKS as a custom application. Each user continues to utilize a variety of typical office applications, but also periodically interacts with SOLIDWORKS. The results are shown in Figure 5, and illustrate that up to 96 VMs can be supported.

It is important to note that scaling was stopped at 96 VMs, even though the CPU utilization is less than 100%. This was necessary because higher consolidation ratios exceeded the View Planner response time threshold and were deemed to have failed the test's performance criteria. This illustrates a key aspect of this paper's approach to scaling: scale-up does not continue until a 100% CPU utilization is reached. Rather, scale-up is only continued as long as the response time meets the View Planner threshold response time, ensuring that users at the peak consolidation ratio see acceptable performance.

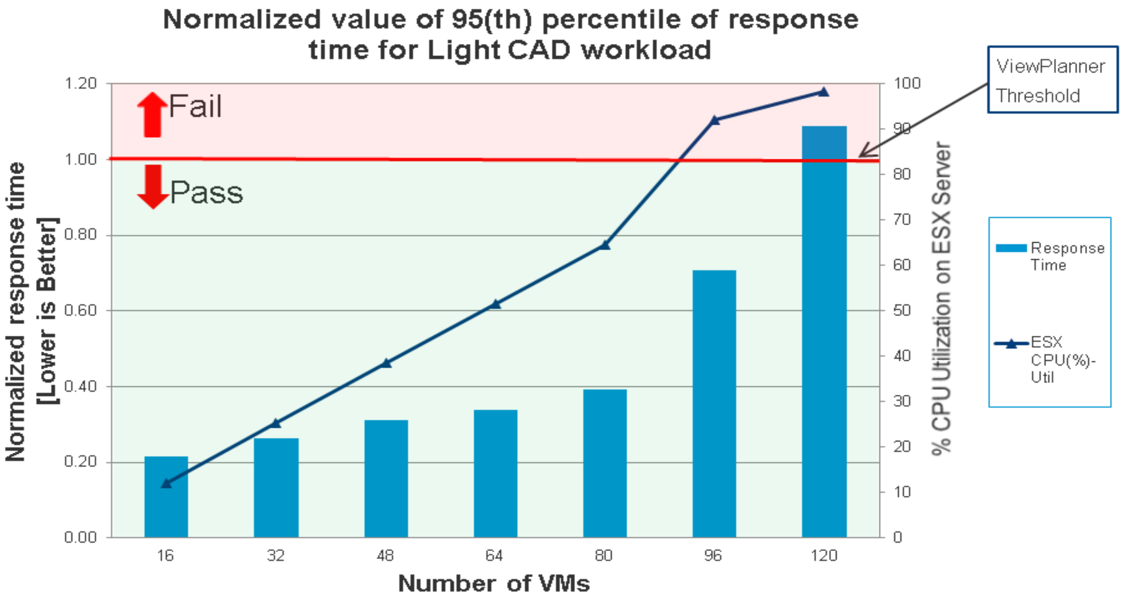


Figure 5. The bar chart presents the normalized values of the 95th percentile of the response time for the light CAD workload as the number of VMs is increased. The corresponding CPU utilization, as measured using esxtop, is shown by the line graph.

CAD Workload 1 (SOLIDWORKS Viewer)

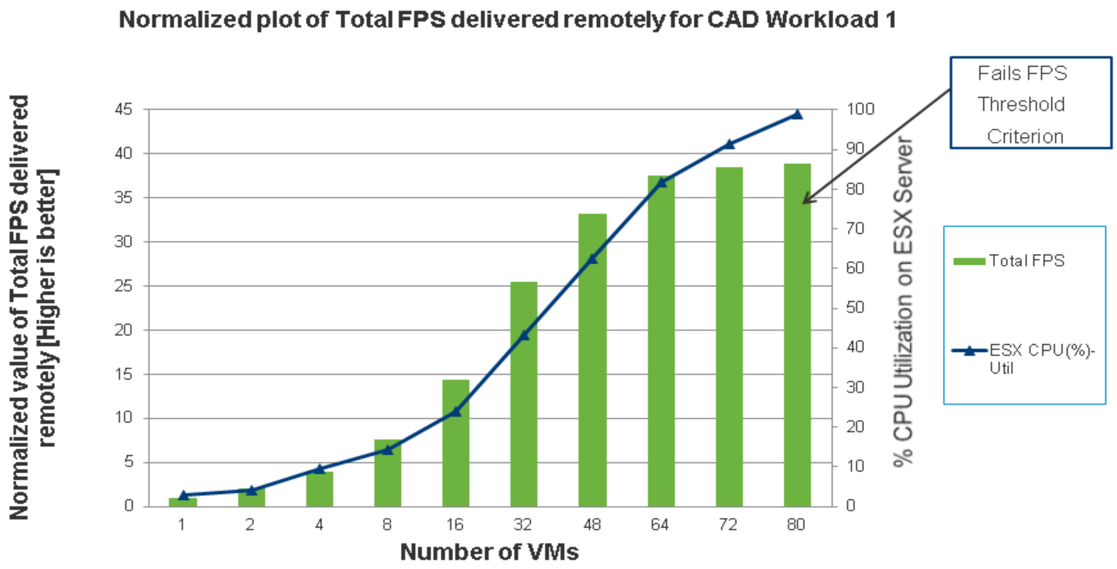


Figure 6(a). The bar chart presents the aggregate frame rate delivered by the system with two NVIDIA K2 cards to the remote clients as the load on the server is increased. The results are normalized and the frame rate observed with just one VM running on the server is defined as the basis for comparison. The corresponding CPU utilization as measured using esxtop is shown by the line graph. GPU utilization was observed to be above 90%.

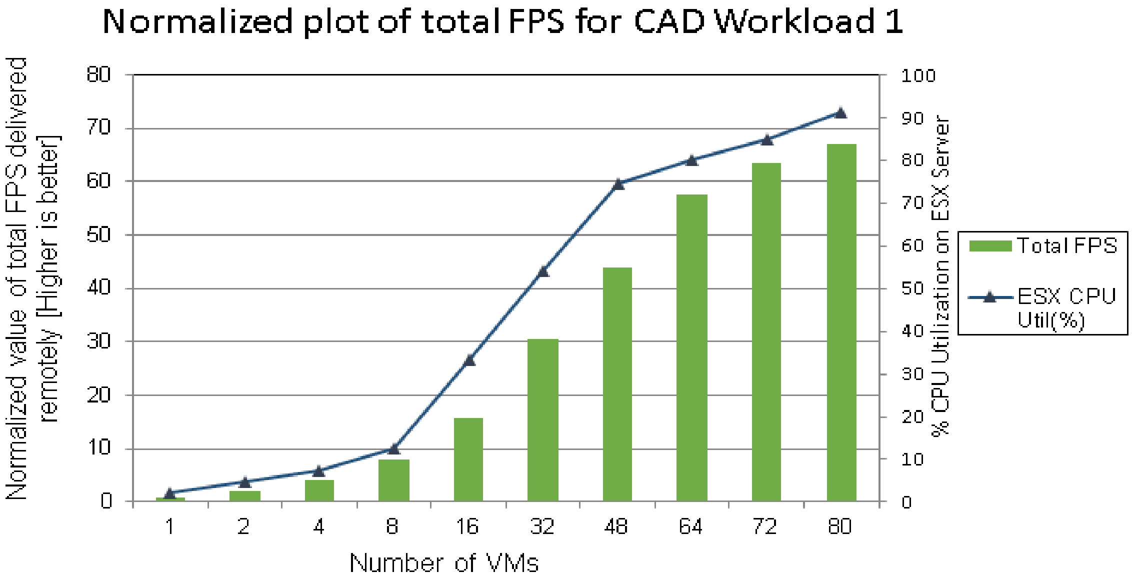


Figure 6(b). The bar chart presents the aggregate frame rate delivered by the system with two NVIDIA K1 cards to the remote clients as the load on the server is increased. The results are normalized and the frame rate observed with just one VM running on the server is defined as the basis for comparison. The corresponding CPU utilization as measured using esxtop is shown by the line graph.

In this workload the office applications used in the prior tests are dropped and attention is focused on the 3D

CAD application. The workload consists of the SOLIDWORKS CAD viewer interacting with the two previously discussed models. The normalized system performance (where performance is defined as the normalized aggregate frame rate delivered by the system to the remote clients) as the number of VMs is increased is presented in Figure 6 and illustrates the scalability of the vSGA stack and Horizon 6: the aggregate performance with 72 desktop VMs is more than 35 times the single VM performance for K2 and the aggregate performance with 80 desktop VMs is more than 70 times the single VM performance for K1.

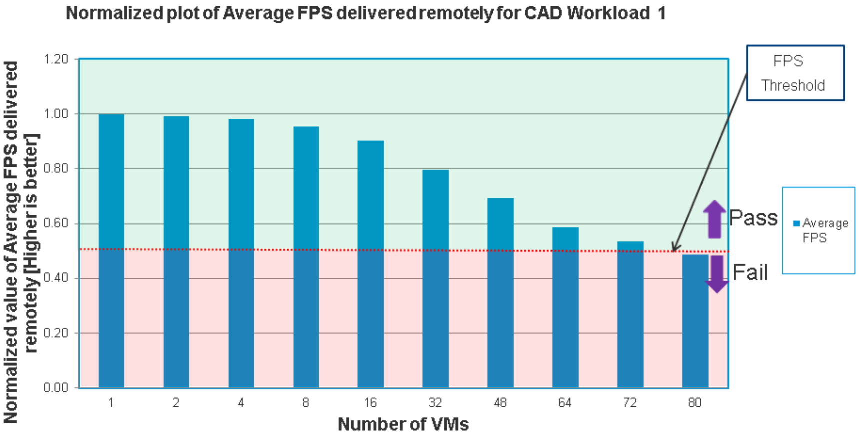


Figure 7(a). The bar chart presents the average frame rates delivered to the Horizon clients as the load on the server with two NVIDIA K2 cards is increased. The results are normalized and the frame-rate observed with just one VM running on the server is defined as the basis for comparison.

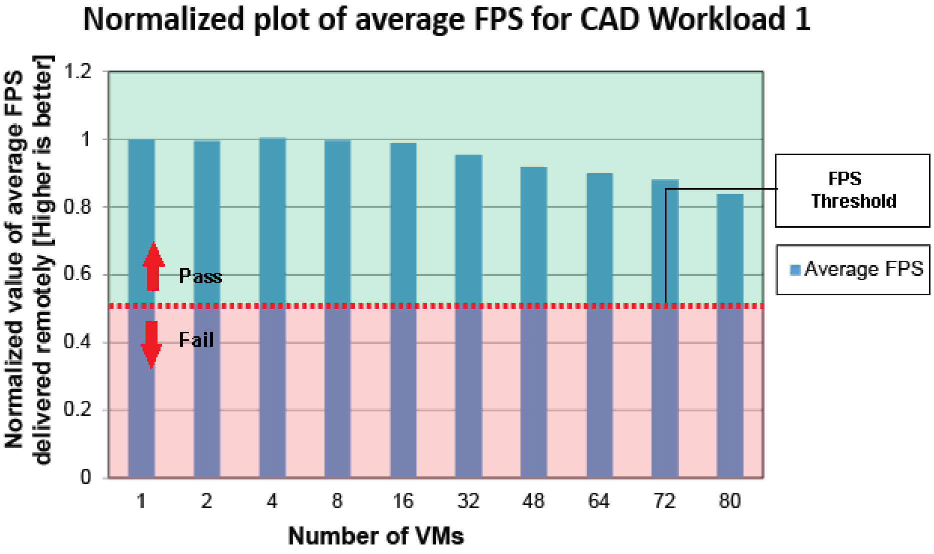


Figure 7(b). The bar chart presents the average frame rates delivered to the Horizon clients as the load on the server with two NVIDIA K1 cards is increased. The results are normalized and the frame-rate observed with just one VM running on the server is defined as the basis for comparison.

Rather than presenting this performance data in terms of the server aggregate, Figure 7 presents the same data on a per VM basis. From the data it is apparent that the per-VM frame-rates are only modestly impacted as the consolidation ratio is steadily increased.

The results in Figure 6 and Figure 7 clearly illustrate the strength of the vSGA solution for 3D workloads: a dual-socket server with 24 cores that might traditionally be provided on a per CAD-user basis is now capable of supporting over 80 CAD users.

CAD Workload 2 (Solid Edge Viewer)

In this workload, the Solid Edge viewer replaces the SOLIDWORKS viewer and runs a single model: a 3-to-1 reducer (as shown at the bottom of Figure 2). 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 8 illustrates the scalability of the vSGA and Horizon View solution; showing the aggregate remotely delivered frame-rates (FPS) as the number of desktop VMs on the server is steadily increased. As the number of VMs is increased from 1 to 72 the aggregate remotely delivered frame-rate increases by 35X for K2 and 30X for K1.

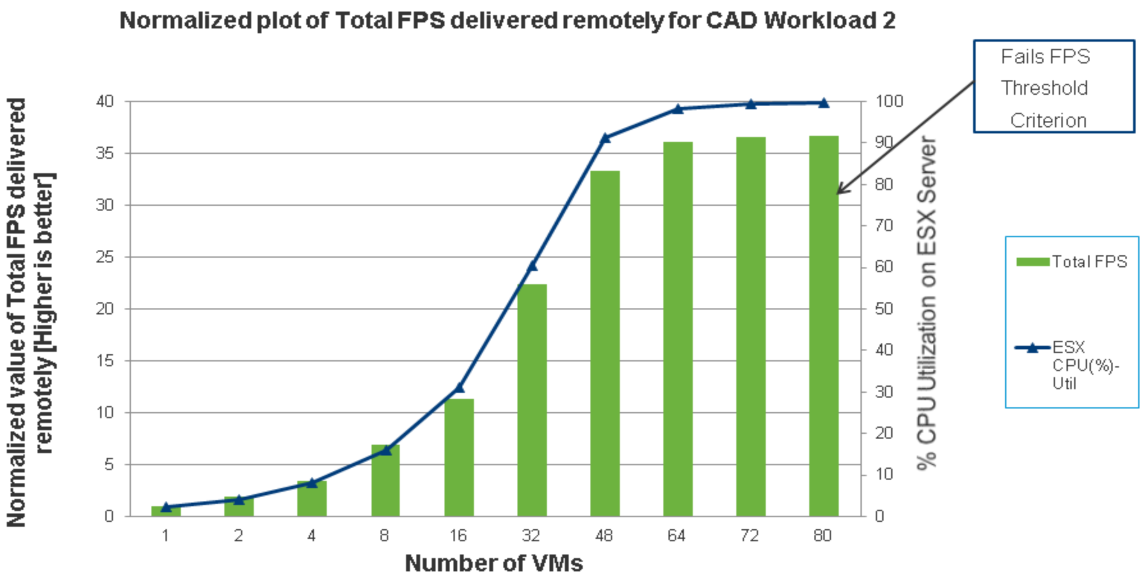


Figure 8(a). This bar chart presents the scalability of the vSGA solution as the load on the server with two K2 cards is increased. The results are normalized and the frame-rate observed with just one VM running on the server is defined as the basis for comparison. The corresponding CPU utilization as measured using esxtop is shown by the line graph. Peak GPU utilization was observed to be about 95%.

Figure 9 presents the same results as Figure 8, but presents the remotely delivered frame-rate data on a per-VM basis. This view of the data highlights that the performance of the individual VMs sees moderate decrease as the number of VMs is scaled.

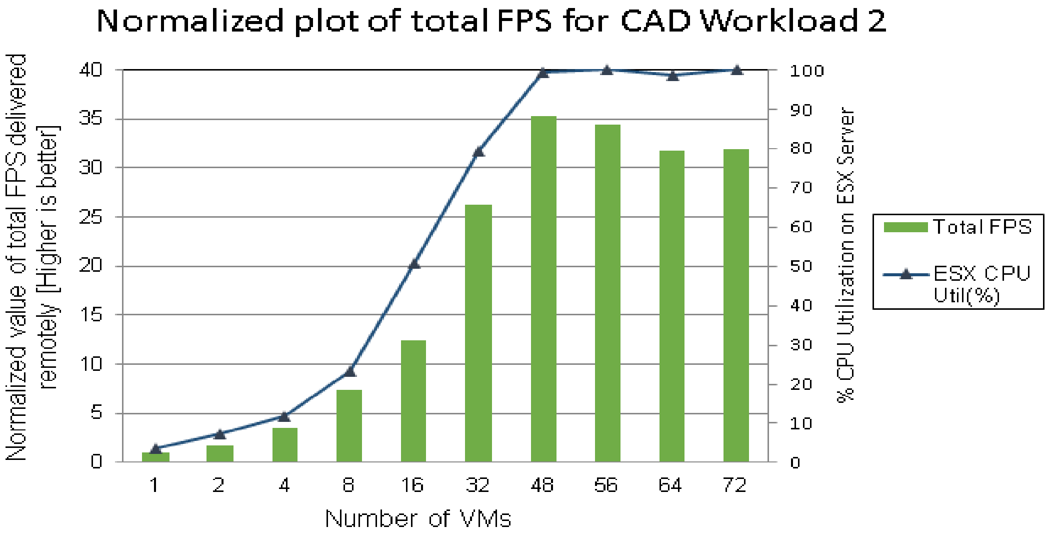


Figure 8(b). This bar chart presents the scalability of the vSGA solution as the load on the server with two K1 cards is increased. The results are normalized and the frame-rate observed with just one VM running on the server is defined as the basis for comparison. The corresponding CPU utilization as measured using esxtop is shown by the line graph. Peak GPU utilization was observed to be about 95%.

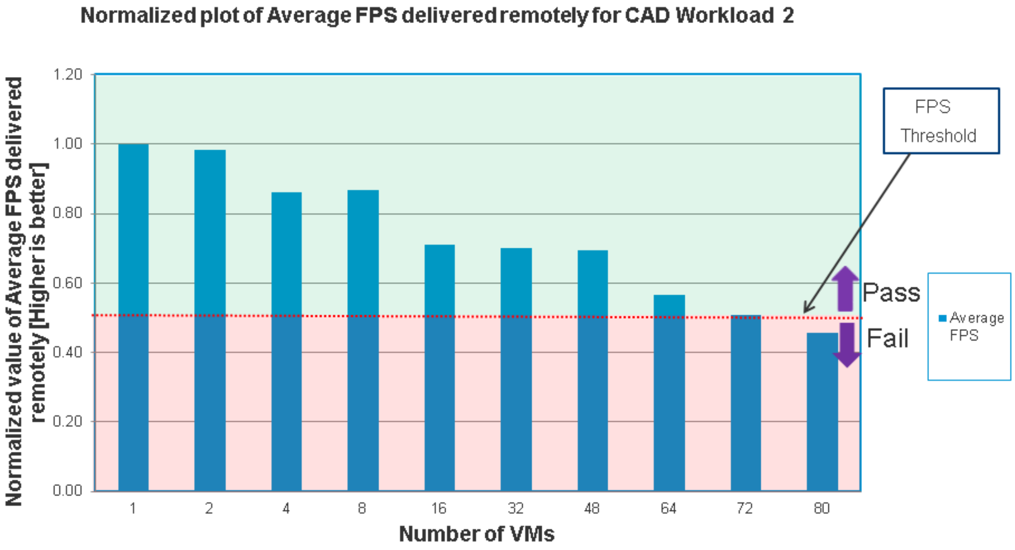


Figure 9(a). The bar chart presents the average per VM remoted frame-rates (FPS) observed with the Solid Edge viewer as the number of VMs on the server with two NVIDIA K2 cards is increased. The results are normalized, with the frame-rate observed with a single VM used as the basis for comparison. In addition, the bar chart is also marked with the associated standard deviation.

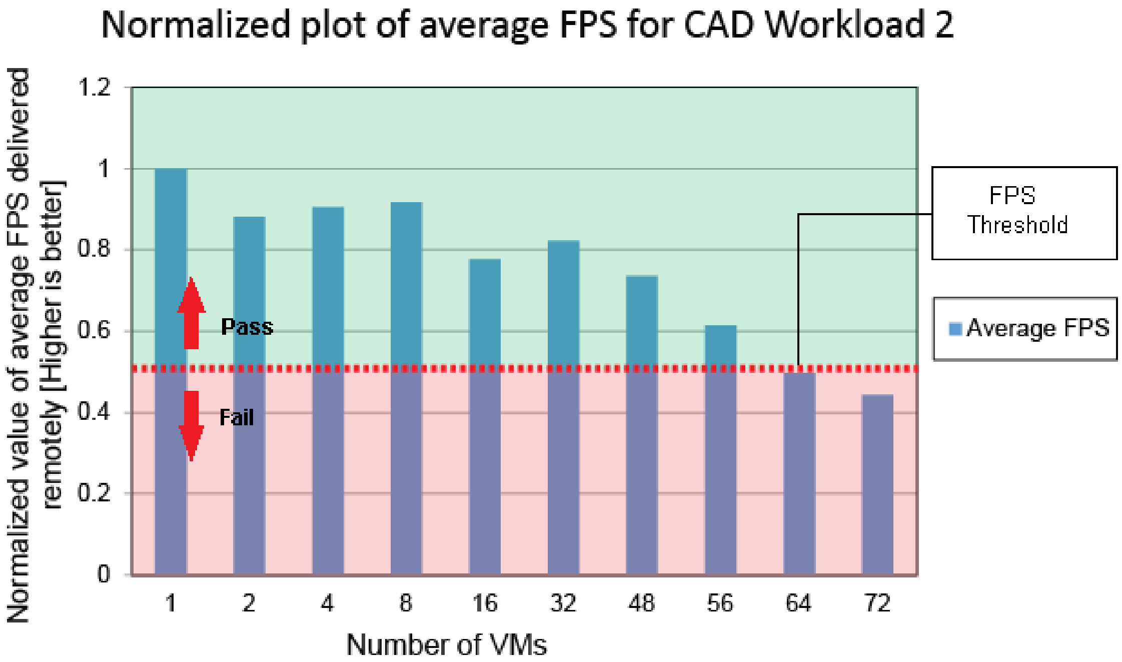


Figure 9(b). The bar chart presents the average per VM remoted frame-rates (FPS) observed with the Solid Edge viewer as the number of VMs on the server with two NVIDIA K1 cards is increased. The results are normalized, with the frame-rate observed with a single VM used as the basis for comparison. In addition, the bar chart is also marked with the associated standard deviation.

Conclusion

This paper presents best practices and performance data for Horizon 6's support for hardware accelerated 3D. The results clearly illustrate the ability of VMware's hardware-backed 3D support to scale efficiently to about 40 3D VMs per physical GPU and, even for performance-intensive 3D workloads, scale efficiently until GPU or CPU resources are exhausted. Specifically, it was demonstrated that using just a mid-range 2-socket x86 server configured with 2 GPUs that significant consolidation of 3D users can be achieved:

- Over 100 users running typical office applications on 3D desktops
- Over 64 users running CAD applications such as SOLIDWORKS or Solid Edge

This clearly shows the benefits of GPU virtualization and the strength of VMware's 3D strategy.

Finally, the paper highlights that very little configuration is required to achieve optimal performance and scalability; vSphere 5.5 and Horizon 6 work well out of the box to efficiently share the CPU and GPU resources between the VMs.

References

[1] VMware Horizon View Documentation. VMware Inc., 2013.
http://www.vmware.com/support/pubs/view_pubs.html

[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, Winter 2012.
http://labs.vmware.com/academic/publications/view-vmtj-winter2012

[3] VMware View Planner Community
http://communities.vmware.com/community/vmtn/servicessoftware/view_planner?view=discussions

[4] VMware View Planner 2.1 Appliance and User Guide
http://communities.vmware.com/docs/DOC-15578

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

[6] View Planner 3.5. http://www.vmware.com/products/view-planner

About the Authors

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

Hari Sivaraman is a Staff Engineer at VMware. He works on 3D rendering performance and on CUDA support on ESXi.

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

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

Acknowledgements

The authors would like to extend their sincere thanks to Julie Brodeur for her comments, feedback, and improvement to the quality of this whitepaper. The authors also thank Warren Ponder, Pat Lee, and Rob Baesman for their comments on the early draft.

VMware, Inc. 3401 Hillview Avenue Palo Alto CA 94304 USA Tel 877-486-9273 Fax 650-427-5001 www.vmware.com