Solution Guide

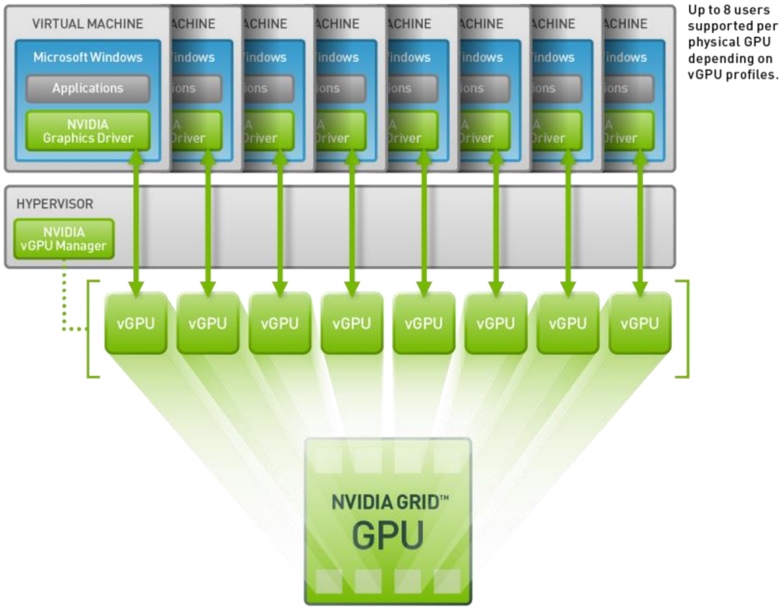
Balancing Graphics Performance, User Density & Concurrency with NVIDIA GRID™ vGPU ™ (Virtual GPU Technology) for Autodesk Revit Power Users

V1.0

Table of Contents

The GRID vGPU Benefit

The inclusion of GRID vGPU™ support in XenDesktop 7.1 allows businesses to leverage the power of NVIDIA's GRID™ technology to create a whole new class of virtual machines designed to provide end users with a rich, interactive graphics experience. By allowing multiple virtual machines to access the power of a single GPU within the virtualization server, enterprises can now maximize the number of users with access to true GPU based graphics acceleration in their virtual machines. Because each physical GPU within the server can be configured with a specific vGPU profile organizations have a great deal of flexibility in how to best configure their server to meet the needs of various types of end users.



Up to 8 VMs can connect to the physical GRID GPU via vGPU profiles controlled by the NVIDIA vGPU Manager.

While the flexibility and power of vGPU system implementations provide improved end user experience and productivity benefits, they also provide server administrators with direct control of GPU resource allocation for multiple users. Administrators can balance user density and performance, maintaining high GPU performance for all users. While user density requirements can vary from installation to installation based on specific application usage, concurrency of usage, vGPU profile characteristics, and hardware variation, it's possible to run standardized benchmarking procedures to establish user density and performance baselines for new vGPU installations.

Understanding GRID vGPU Profiles

Within any given enterprise the needs of individual users varies widely, a one size fits all approach to graphics virtualization doesn't take these differences into account. One of the key benefits of NVIDIA GRID vGPU is the flexibility to utilize various vGPU profiles designed to serve the needs of different classes of end users. While the needs of end users can be quite diverse, for simplicity we can group them into the following categories: Knowledge Workers, Designers and Power Users.

For **knowledge workers** key areas of importance include office productivity applications, a rich web experience, and fluid video playback. Graphically knowledge workers have the least graphics demands, but they expect a similarly smooth, fluid experience that exists natively on today's graphic accelerated devices such as desktop PCs, notebooks, tablets and smart phones.

Power Users are those users with the need to run more demanding office applications; examples include office productivity software, image editing software like Adobe Photoshop, mainstream CAD software like Autodesk Revit and product lifecycle management (PLM) applications. These applications are more demanding and require additional graphics resources with full support for APIs such as OpenGL and Direct3D.

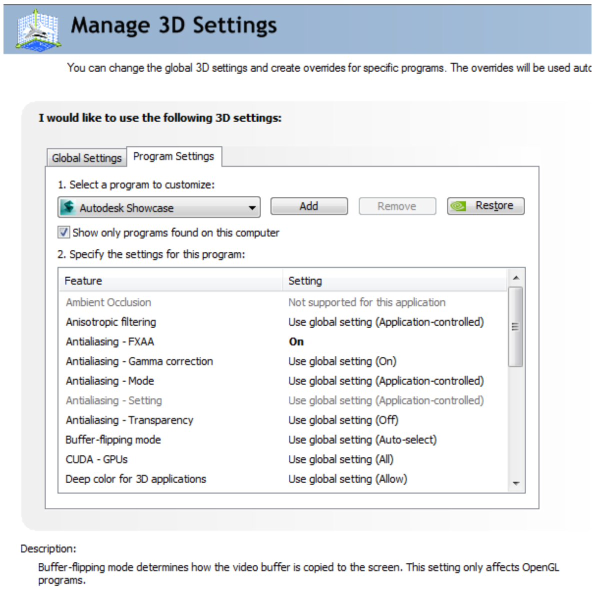
Designers are those users within an organization running demanding professional applications such as high end CAD software and professional digital content creation (DCC) tools. Examples include Autodesk Inventor, PTC Creo, Autodesk Revit and Adobe Premiere. Historically designers have utilized desktop workstations and have been a difficult group to incorporate into virtual deployments due to the need for high end graphics, and the certification requirements of professional CAD and DCC software.

The various NVIDIA GRID vGPU profiles are designed to serve the needs of these three categories of users:

vGPU Profile	GRID Card	Use Case	Framebuffer (MB)	Maximum VM's Per GPU	Maximum VM's Per Card
GRID K100	GRID K1	Knowledge Worker	256	8	32
GRID K140Q	GRID K1	Power User	1024	4	16
GRID K200	GRID K2	Knowledge Worker	256	8	16
GRID K240Q	GRID K2	Designer / Power User	1024	4	8
GRID K260Q	GRID K2	Designer / Power User	2048	2	4

The GPU profiles ending in Q are certified graphic solutions for professional applications such as Autodesk Inventor 2014 and PTC Creo, undergoing the same rigorous application certification testing as NVIDIA's Quadro workstation products.

Each GPU within a system must be configured to provide a single vGPU profile, however separate GPU's on the same GRID board can each be configured separately. For example a single K2 board could be configured to serve eight K200 enabled VM's on one GPU and two K260Q enabled VM's on the other GPU.
The key to efficient utilization of a system's GRID resources requires understanding the correct end user workload to properly configure the installed GRID cards with the ideal vGPU profiles maximizing both end user productivity and vGPU user density.



Manage 3D Settings

You can change the global 3D settings and create overrides for specific programs. The overrides will be used autc

I would like to use the following 3D settings:

Global Settings | Program Settings

1. Select a program to customize:

Autodesk Showcase

☑ Show only programs found on this computer

2. Specify the settings for this program:

Feature	Setting
Ambient Occlusion	Not supported for this application
Anisotropic filtering	Use global setting (Application-controlled)
Antialiasing - FXAA	**On**
Antialiasing - Gamma correction	Use global setting (On)
Antialiasing - Mode	Use global setting (Application-controlled)
Antialiasing - Setting	Use global setting (Application-controlled)
Antialiasing - Transparency	Use global setting (Off)
Buffer-flipping mode	Use global setting (Auto-select)
CUDA - GPUs	Use global setting (All)
Deep color for 3D applications	Use global setting (Allow)

Description:

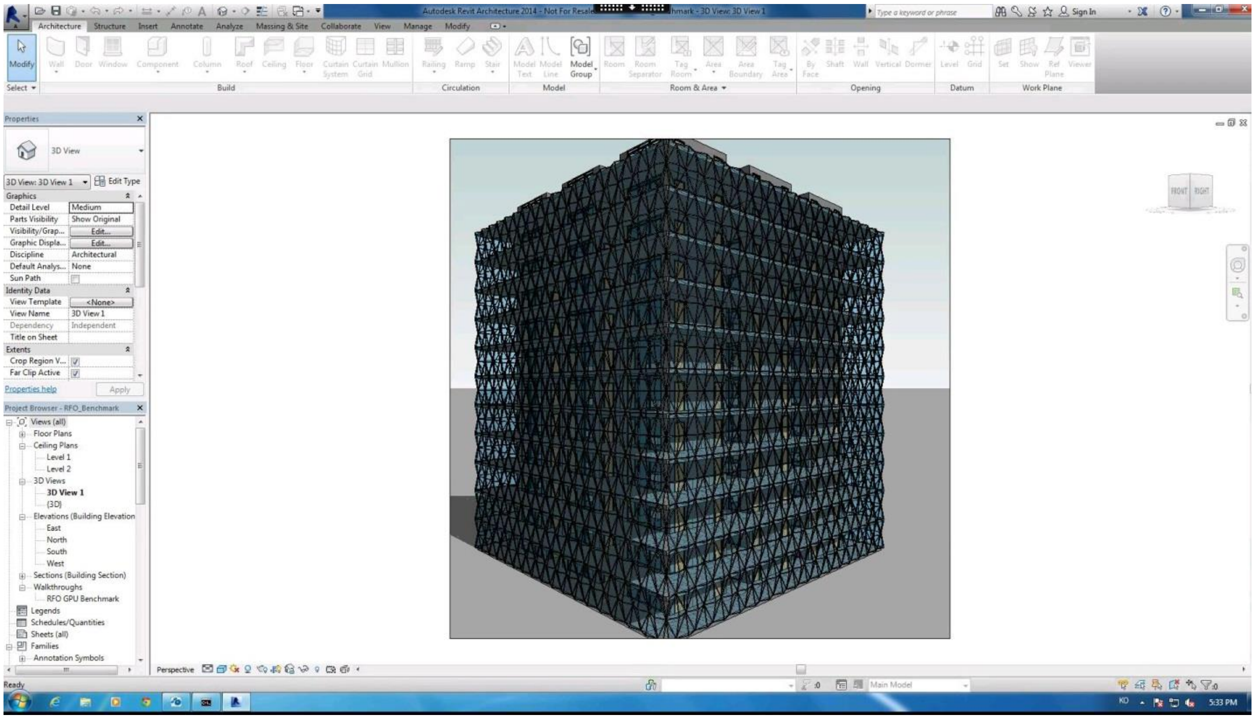
Buffer-flipping mode determines how the video buffer is copied to the screen. This setting only affects OpenGL programs.

The vGPU profiles with the "Q" suffix (K140Q, K240Q and K260Q), offer additional benefits not available in the non-Q profiles, the primary of which is that Q based vGPU profiles will be certified for professional applications. These profiles offer additional support for professional applications by optimizing the graphics driver settings for each application using NVIDIA's **Application Configuration Engine** (ACE), ACE offers dedicated profiles for most professional workstation applications, once ACE detects the launch of a supported application it verifies that the driver is optimally tuned for the best user experience in the application.

Benchmarking as a Proxy for Real World Workflows

In order to provide data that offers a positive correlation to the workloads we can expect to see in actual use, benchmarking test case should serve as a reasonable proxy for the type of work we want to measure. A benchmark test workload will be different based on the end user category we are looking to characterize. For knowledge worker workloads a reasonable benchmark is the Windows Experience Index, and for Power Users we can use the Revit benchmark for Autodesk Revit. The SPEC Viewperf benchmark is a good proxy for Designer use cases.

To illustrate how we can use benchmark testing to help determine the correct ratio between total user density and workload performance we'll look at a Power User workload using the Revit benchmark, which tests performance within Autodesk Revit 2014. The benchmark tests various aspects of Revit performance by running through a series of common workloads used in the creation of a Revit project. These workloads include viewport rotation and viewport refresh using realistic and hidden line visual styles. These areas have been identified in particular as pain points within the average users Revit workflow. The benchmark creates a detailed model and then automates interacting with this model within the application viewports in real-time.



3D View: 3D View 1

Graphics	
Detail Level	Medium
Parts Visibility	Show Original
Visibility/Grap...	Edit...
Graphic Displa...	Edit...
Discipline	Architectural
Default Analys...	None
Sun Path	□
Identity Data	
View Template	<None>
View Name	3D View 1
Dependency	Independent
Title on Sheet	
Extents	
Crop Region V...	☑
Far Clip Active	☑

The Revit benchmark is an excellent proxy for end user workloads, it is designed to test the creation of an actual real world model and test performance using various graphic display styles and return a benchmark score which isolates the various performance categories. Because the benchmark runs without user interaction once started it is an ideal candidate for multi-instance testing. As an industry standard benchmark, it has the benefit of being a credible test case, and since the benchmark shows positive scaling with higher end GPU's it allows us to test various vGPU profiles to understand how profile selection affects both performance and density.

Methodology

By utilizing test automation scripting tools, we can automate launching the benchmark on the target VM's. We can then automate launching the VM's so that the benchmark is running on the target number of VM's concurrently. Starting with a single active user per physical GPU, the benchmark is launched by the client VM and the results of the test are recorded. This same procedure is repeated by simultaneously launching the benchmark on additional VM's and continuing to repeat these steps until the maximum number of vGPU accelerated VMs per GRID card (K1 or K2) is reached for that particular vGPU profile.

Fully Engaged Graphics Workloads?

When running benchmark tests, we need to determine whether our test nodes should be fully engaged with a graphics load or not. In typical real-world configurations the number of provisioned VM's actively engaged in performing graphically intensive tasks will vary based on need within the enterprise environment. While possible, it is highly unlikely that every single provisioned VM is going to be under a high demand workload at any given moment in time.

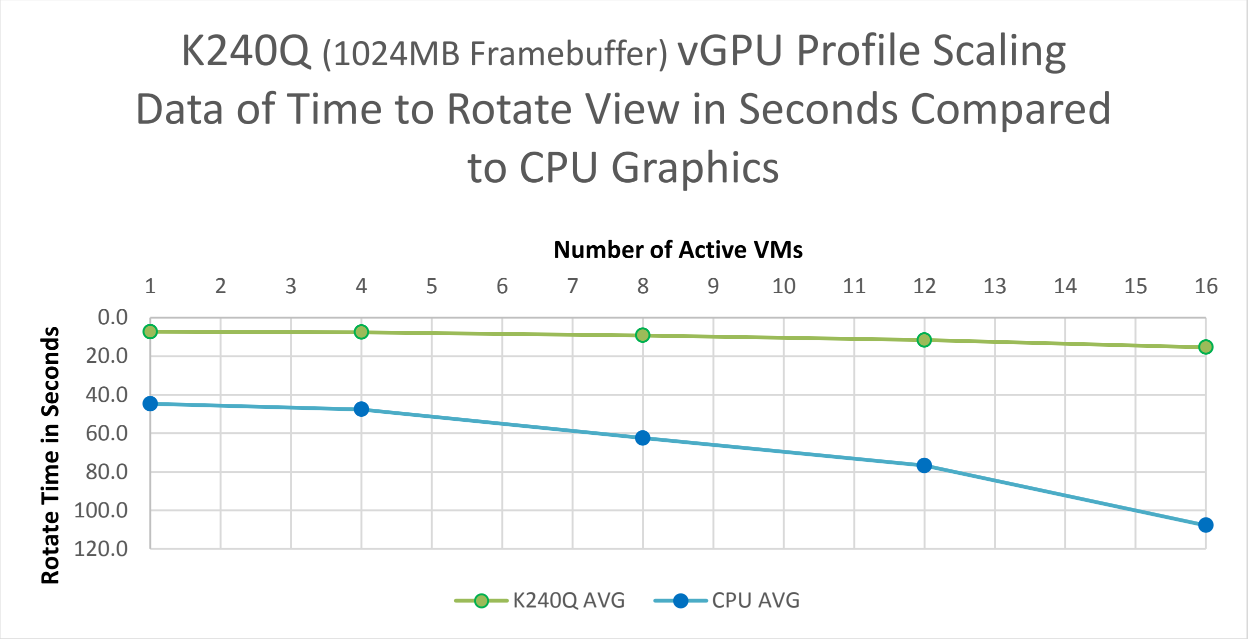
In setting up our benchmarking framework we have elected to utilize a scenario that assumes that every available node is fully engaged. While such heavy loading is unlikely to occur in a real world environment, it allows us to use a “worst case scenario” to plot our density vs. performance data.

Analyzing the Performance Data to Understand How User Density Affects Overall Performance

To analyze the benchmark result data it’s important to understand that we are less interested in individual performance results than we are in looking for the relationship between overall performance and total user load. By identifying trends within the results where performance shows a rapid falloff we can begin to make an educated determination about the maximum number of Revit users we can support per server. Because we are most interested in maintaining interactivity within the viewport, we’ll focus on the benchmark results from the **Rotate View** test. To measure scalability we take the sum of the individual result scores from each VM and total them. The total is then divided by the total number of active VM’s to obtain an **Average Score Per VM.** In determining the impacts of density on overall benchmarking performance we plot the benchmark as seen in the graphs below. For each plot we record the average results for each portion of the benchmark score result, and indicate the percentage drop in performance compared to the same profile with a single active VM. Because Revit is an application which certifies professional graphics for use with the application, we can focus on the professional “Q” profiles, 140Q , 240Q and 260Q which are certified options for Revit.

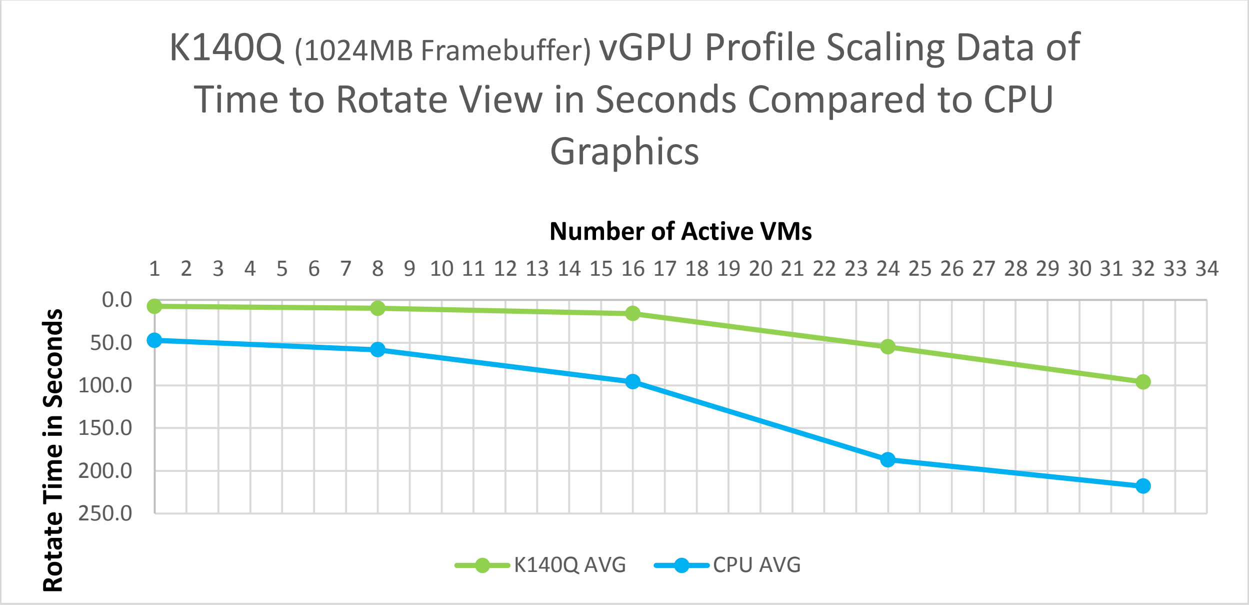
All our testing is done with 2 x GRID boards installed in the server (*2x* K1 or *2x* K2).

In Example 1 below we analyze the data for the K240Q vGPU profile, one of the professional profiles available on the K2 GRID board. The K240Q profile provide 1028MB of framebuffer on the virtual GPU. The performance trend for the K240Q profile show a performance falloff of 109% between a single fully engaged K240Q VM and the maximum number of K240Q fully engaged VM’s supported on the server (16).We can see the superior performance offered by vGPU in the Revit benchmark when running the maximum number of VMs on a dual K2 boards (16), completes the benchmark rotation test 192% faster than a server running a **single** VM instance of the benchmark using CPU emulated graphics and is 614% faster than CPU emulated graphics running the same number of active VMs (16). As the number of active VM’s increases on the server, the results show a performance falloff of 109% between a single fully engaged K240Q VM and the maximum number of K240Q fully engaged VM’s supported on the server (16).



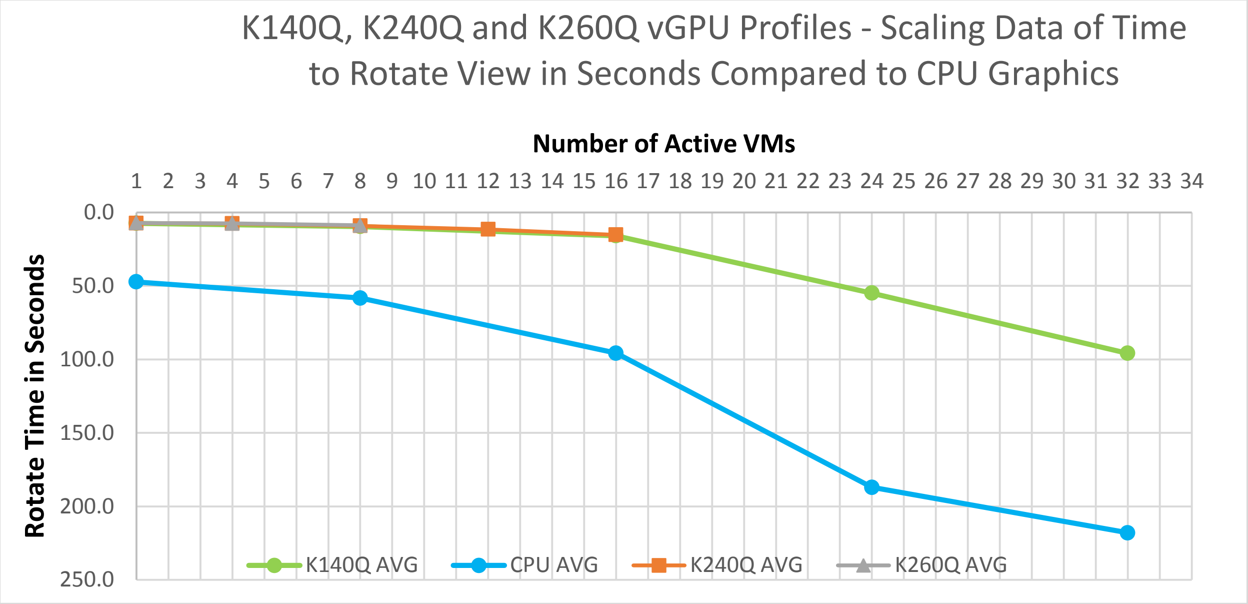
Example 1 – Dual K2 boards allocated with K240Q vGPU profile (1024MB Framebuffer), each K2 board can support up to 8 K240Q vGPU accelerated VMs.

In Example 2 below is the Revit performance profile for the K140Q the professional profile for the K1 GRID board. The K140Q profile is configured with 1024MB of framebuffer per accelerated VM, the same as the K240Q. On a single K1 GRID board the performance profile is extremely similar between the K140Q and the K240Q profiles up to 8 active VMs, which is the maximum number of VMs supported on the K240Q. Moving beyond 8 VM's we see that although the average benchmark scores continue to decline the decline continues at a gradual pace until we get beyond 16 active VM's. Beyond 16 active VM's we see a much more rapid falloff in terms of performance until at around 24 active VM's we see a performance level that falls below the performance of a single CPU emulated graphics VM for the first time, although performance is still significantly better than a CPU emulated graphics configuration running a matching number of active VM's.



Example 2 Dual K1 boards allocated with K140Q vGPU profile (1024MB Framebuffer), each K1 board can support up to 16 K140Q vGPU accelerated VMs for a total of 32 VMs in the tested configuration.

Example 3 below shows the combined performance profiles for both the K2 GRID based K240Q and K260Q profiles and the GRID K1 based K140Q profile compared to CPU emulated graphics showing the results of the Revit benchmark rotate view portion of the test. The performance data for all three GRID profiles are virtually identical. It's worth noting that the trend of performance falloff is similar between the vGPU results and the CPU graphics results. The similarity in falloff is likely an indication that the falloff represents a lack of enough system resources on the server as the number of fully engaged VMs increases past at certain point (for our hardware configuration that point is seen around 16 VMs). The results show that regardless of profile used vGPU offers a significant performance increase over CPU emulated graphics under the same workload.



Example 3 – K260Q, K240Q, and K140Q vGPU profiles show very similar performance and falloff curve matches the CPU falloff curve indicating that system resources are likely the limiting factor.

Board	Profile	Maximum VMs per Board	Recommended range of VM's per server as configured for test.
K1	K140Q	16	16-24 (2x GRID K1)
K2	K240Q	8	16 (2x GRID K2)
K2	K260Q	4	8 (2x GRID K2)

Table 1 – Maximum and recommended VM's per GRID board by profile

Server Configuration

Dell R720
Intel® Xeon® CPU E5-2670 2.6GHz, Dual Socket (16 Physical CPU, 32 vCPU with HT)
Memory 384GB
XenServer 6.2 + SP1

Virtual Machine Configuration

VM Vcpu : 4 Virtual CPU
Memory : 5GB
XenDesktop 7.1 RTM HDX 3D Pro
Revit 2014
Revit Benchmark

NVIDIA Driver: 332.07
Guest Driver: 331.30

Additional NVIDIA GRID Resources

Website – www.nvidia.com/vdi
NVIDIA GRID Forums - https://gridforums.nvidia.com
Certified Platform List – www.nvidia.com/wheretobuy
ISV Application Certification – www.nvidia.com/gridcertifications
GRID YouTube Playlist – www.tinyurl.com/gridvideos

Have issues or questions? Contact us through the NVIDIA GRID Forums or via Twitter @NVIDIAGRID