Solution Guide

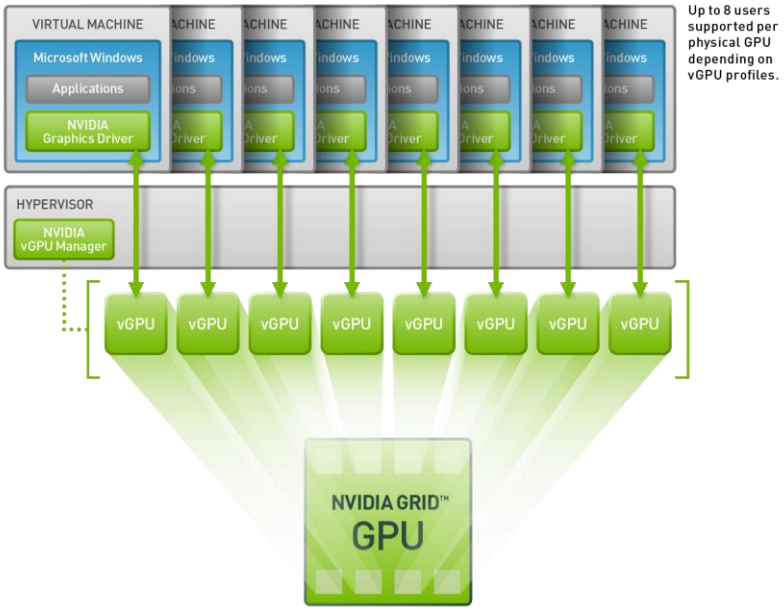
Balancing Graphics Performance, User Density & Concurrency with NVIDIA GRID™ Virtual GPU Technology (vGPU™) for Autodesk AutoCAD Power Users

V1.0

Table of Contents

The GRID vGPU Benefit

The inclusion of **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 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 AutoCAD 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.

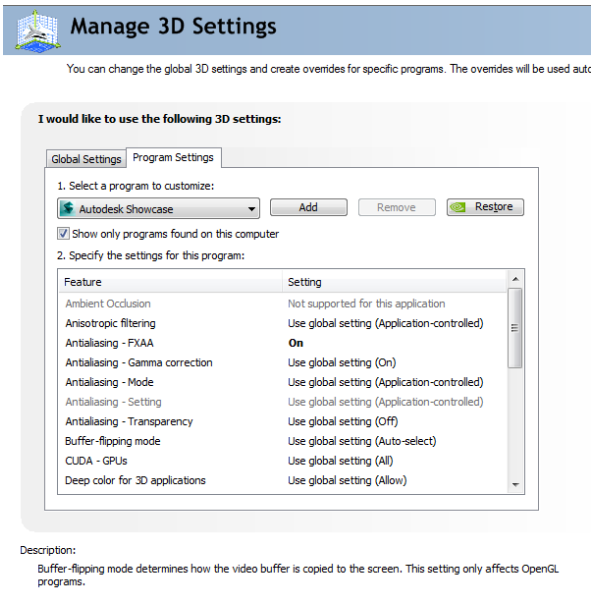
Designers are those users within an organization running demanding professional applications such as high end CAD software and professional digital content creation (DCC) tools. Examples include Autodesk Inventor, PTC Creo, Autodesk Revit and Adobe Premiere. Historically designers have utilized desktop 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.

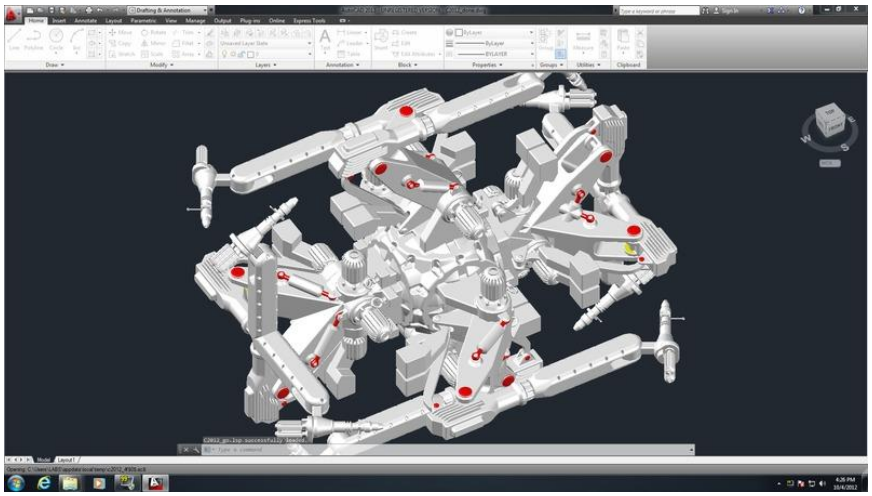


The vGPU profiles with the "Q" suffix (K140Q, K240Q and K260Q), offer additional benefits not available in the non-Q profiles, the primary of which is that Q based vGPU profiles will be certified for professional applications. These profiles offer additional support for professional applications by optimizing 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 CADALYST benchmark for AutoCAD. 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 CADALYST benchmark, which tests performance within Autodesk AutoCAD 2014. The benchmark tests various aspects of AutoCAD performance by loading a variety of models an interacting with them within the application viewports in real-time.



CADALYST offers many advantages for use as a proxy for end user workloads, it is designed to test actual real world models using various graphic display styles and return a benchmark score dedicated to graphics performance. 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?

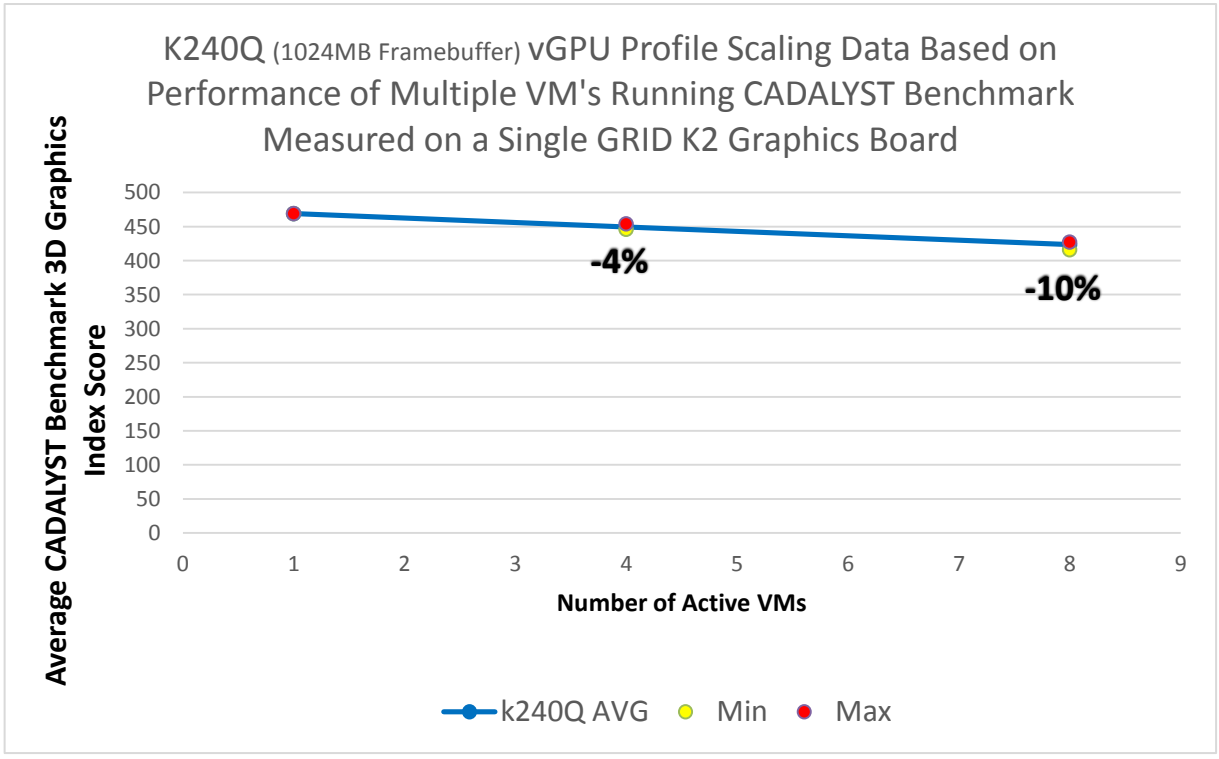
When running benchmark tests, we need to determine whether our test nodes should be fully engaged with a graphics load or not. In typical real-world configurations the number of provisioned VM's actively engaged in performing graphically intensive tasks will vary based on need within the enterprise environment. 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

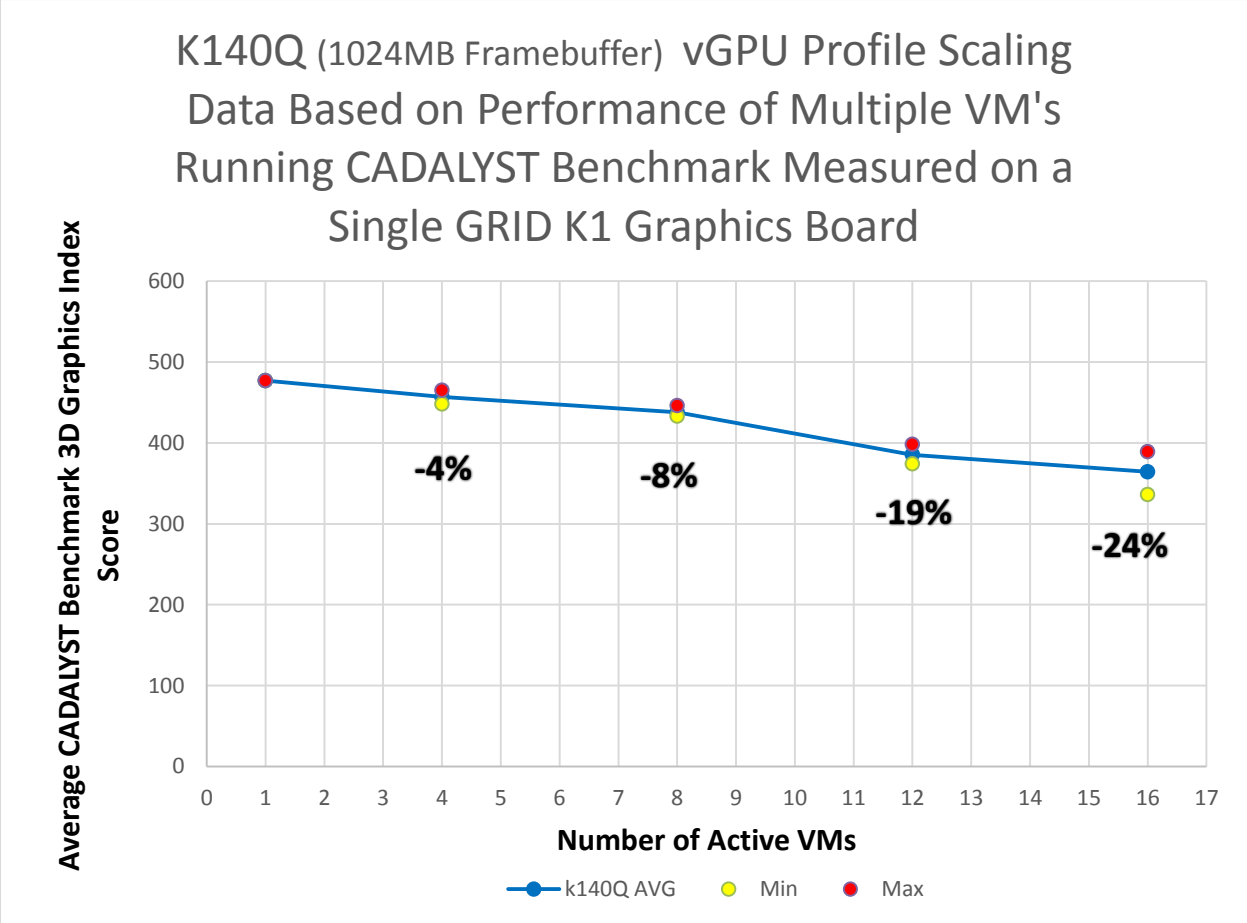
To analyze the benchmark result data we take the sum of the CADALYST 3D result score from each VM and total them. The total is then divided by the total number of active VM's to obtain an **Average Score Per VM.** In determining the impacts of density on overall benchmarking performance we plot the benchmark as seen in the graphs below. For each plot we record the average CADALYST 3D score result, and indicate the percentage drop in performance compared to the same profile with a single active VM. In general the results below show that vGPU profiles with are targeted for Power Users and Designers, experience less performance falloff than profiles which are intended for use by Knowledge workers.

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 performance trend for the K240Q profile shows that performance in the CADALYST benchmark when running the maximum number of VMs on a single K2 board (8), is only 10 percent slower than the performance of a single K240Q accelerated VM active on the server. Overall, adding additional CADALYST workloads to the system shows that performance is minimally impacted by scaling up the number of users on the system. While the blue line on the chart shows the average CADALYST benchmark scores as measured across all active VMs, the Minimum and Maximum scores are also plotted on the graph. For the K240Q profile we see extremely little deviation between the average score and the min and max.



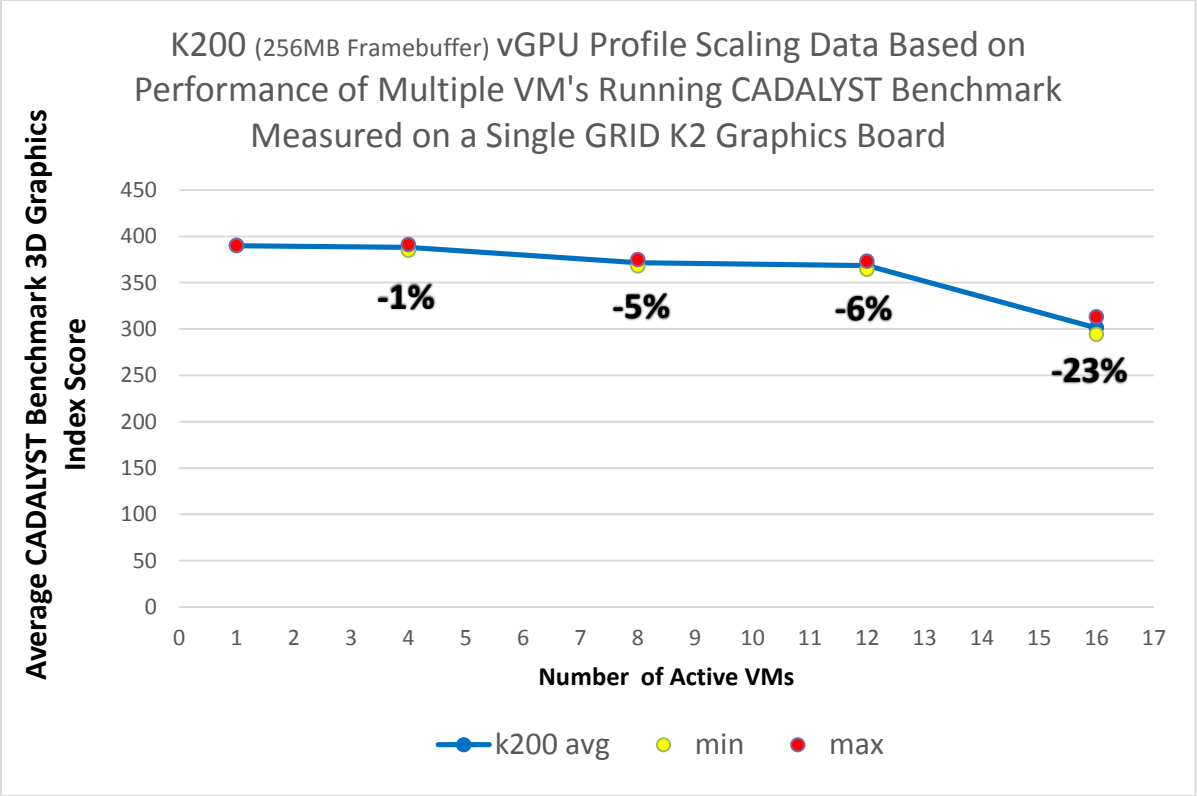
Example 1 – Single K2 board allocated with K240Q vGPU profile (1024MB Framebuffer), each K1 board can support up to 8 K240Q vGPU accelerated VMs.

In Example 2 below is the CADALYST 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 even with the maximum number of K140Q profiles running average scores only drop by 25% compared to a single K140Q accelerated VM running the benchmark on the server. The deviation between the average score and the min/max starts low but increase as more active VMs are added.



Example 2 Single K1 board allocated with K140Q vGPU profile (1024MB Framebuffer), each K1 board can support up to 16 K140Q vGPU accelerated VMs.

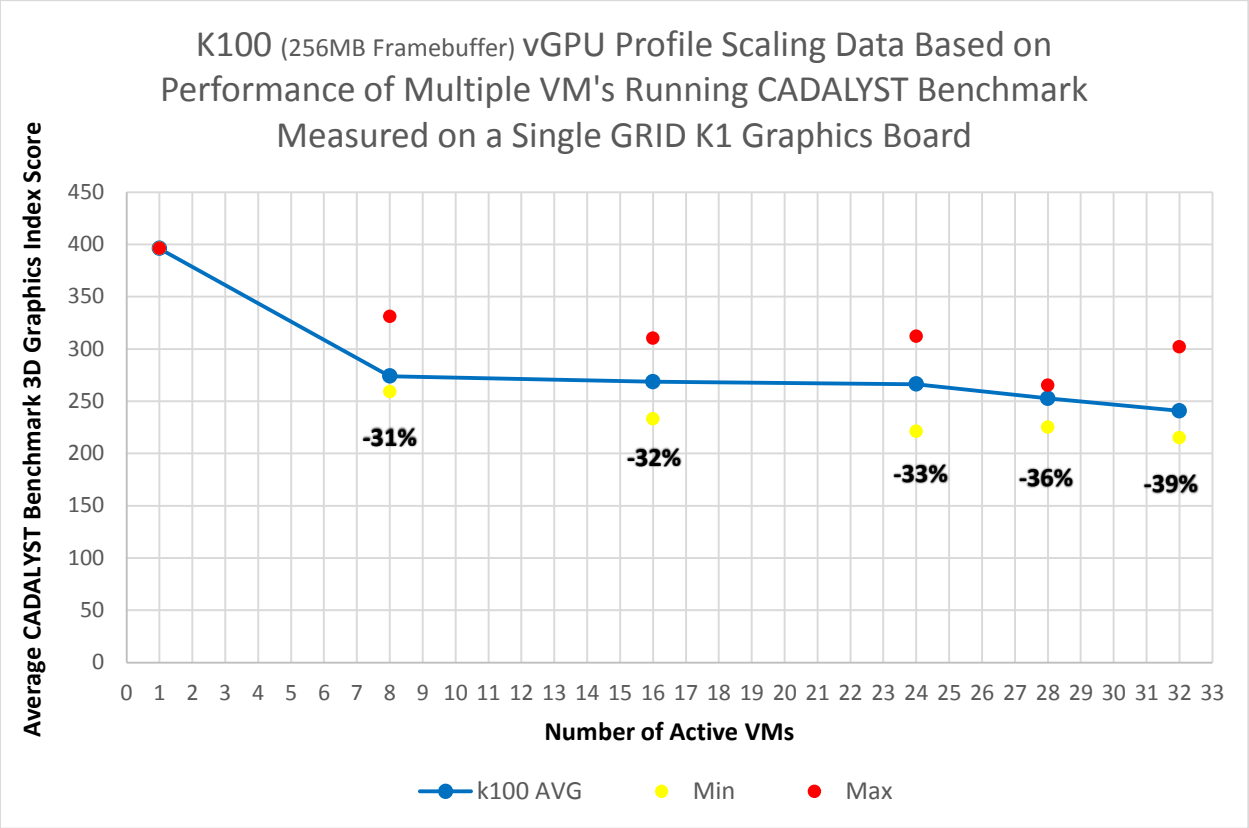
Example 3 below shows the performance profile for a single K2 GRID board using K200 vGPUs. The chart shows that adding additional K200 accelerated VMs has minimal impact on performance up to VMs. At the maximum number of VMs supported on a single K2 board (16) there is a more pronounced falloff between 12 and 16 active VMs. Although the total percentage of performance drop at 16 VMs is similar between K200 and K140Q, the professional based K140Q profile offers higher average performance with the maximum number of VMs actively running the benchmark.



Example 3 – Single K2 board allocated with K200 vGPU profile (256MB Framebuffer), each K2 board can support up to 16 K200 vGPU accelerated VMs.

In Example 4 we see the performance profile for a single GRID K1 board configured with K100 vGPUs. The performance trend for the K100 profile shows that when loaded with eight actively engaged VMs (25% of a K1 board's maximum capacity when allocated as a K100 vGPU profile), there is a noticeable 31% percent drop in average performance. However, after the initial performance drop, adding additional engaged VM's up to the maximum of 32 only results in minimal additional falloff. This indicates that if application performance is acceptable with 8 engaged users, that adding more users to the system with similar workloads shouldn't negatively affect the overall system performance in a noticeable manner.

In addition to the average score for all active VM's, the graph also indicates the minimum and maximum scores recorded by all VM's generating benchmark results. We can see that while the average performance drop between a single active VM and the maximum of 32 supported by a single K1 board is 39%, in some cases the performance drop was as little as 24%



Example 4 – Single K1 board allocated with K100 vGPU profile (256MB Framebuffer), each K1 board can support up to 32 K100 vGPU accelerated VMs.

Board	Profile	Maximum VMs per Board	Recommended range of VM's per Board
K1	K100	32	20 - 32
K1	K140Q	16	12 - 16
K2	K200	16	10 - 16
K2	K240Q	8	7 - 8
K2	K260Q	4	3 - 4

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

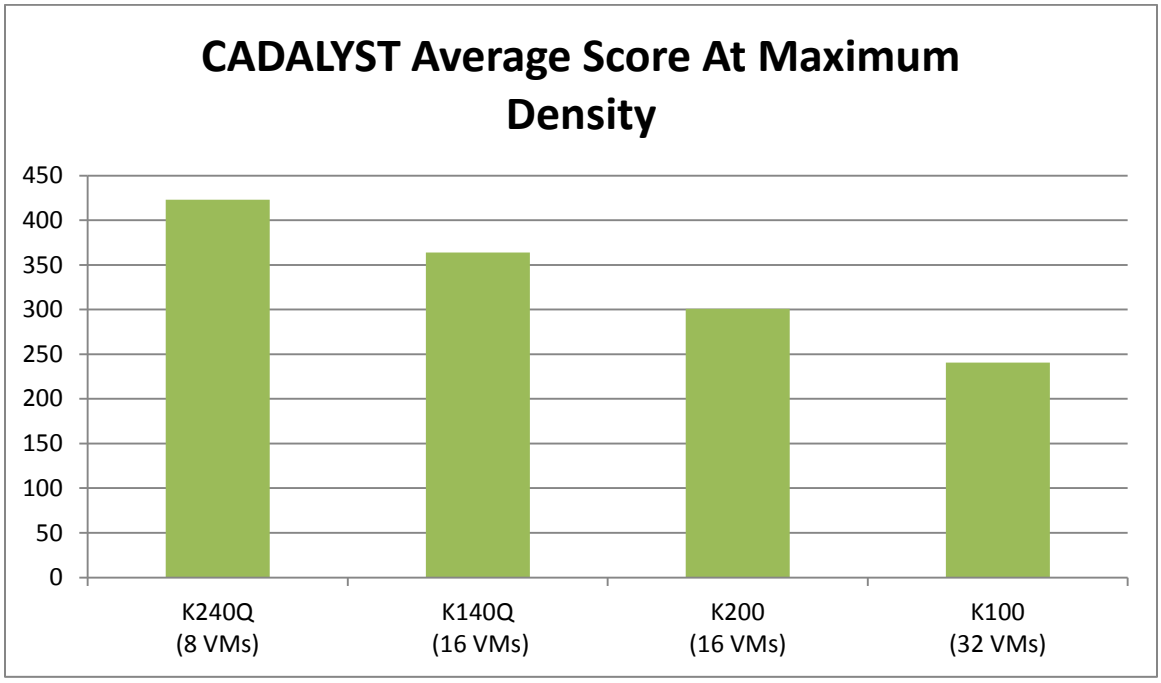


Chart 1 – Average 3D benchmark score at recommended densities.

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 Tech Preview Build 74074c

Virtual Machine Configuration

VM Vcpu : 4 Virtual CPU
Memory : 11GB
XenDesktop 7.1 RTM HDX 3D Pro

AutoCAD 2014
CADALYST C2012 Benchmark

NVIDIA Driver: vGPU Manager : 331.24

Guest driver : 331.82

Additional GRID Resources

Website – www.nvidia.com/vdi

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 setting up or viewing demos? Contact the GRID demo team via email at demogrid@nvidia.com or @NVIDIAGRID on Twitter.