NVIDIA GRID:
Deployment Best Practices for the Digital Workplace

Whitepaper

Author: Nachiket Karmarkar, NVIDIA Performance Engineer

WP-09455-001_v01 | April 2019

Table of Contents

NVIDIA GRID for the New Digital Worker in Virtual Desktop Environments

To provide the right level of user experience for your users on virtual desktops, it's important to provide them with the right amount of resources based on their workloads. There are many factors to consider when sizing a virtual desktop environment including CPU, RAM, storage, network, and more. The purpose of this whitepaper is to provide guidance on how to assign the appropriate amount of frame buffer for the "New Digital Worker" on NVIDIA GRID® environments. The new term digital worker refers to the new way we work in the tech-savvy, digital age. Across industries, the workplace has become more visual - video, web conferencing, multi-monitors, dynamic browsers, 3D features in everyday applications, and more, are the new normal. Even Windows 10 is more graphics intensive than Windows 7, requiring 50% more graphics calls[1].

With the upcoming end of life (EOL) of Windows 7, enterprises are migrating to Windows 10 and may need to refresh their data center infrastructure as well to support the new digital worker. The latest operating system from Microsoft, Windows 10, was designed to deliver improved user experience on both PCs and mobile devices. Windows 10 was developed with hardware acceleration in mind. Hardware acceleration means using a computer's hardware to do certain tasks and functions faster than is possible using software. Moving graphics rendering from the CPU to the GPU allows better graphics performance. Because Windows 10 is fully accelerated using a graphics processing unit (GPU) by default or is left to emulate a GPU with software rendering when one is unavailable, the operating system's full potential can't be realized without graphics acceleration.

It has become tougher to disable Windows 10 features during the initial releases, a common practice for many VDI environments. For the later releases, while there are options to disable visual effects and animations to decrease resource utilization, you cannot completely turn off the Metro/Modern UI. The practice of stripping down the system and removing apps and

[1] Lakeside Software, comparing the percent of time the OS is consuming GPU (DirectX or OpenGL) from Windows 7 to Windows 10 in 2018 (builds 1803 and 1709).

services is frustrating to end-users who expect their Windows virtual PC at work to function the same as their Windows PC at home.

And while Windows 10 has a great look and feel, it also introduces more frequent OS updates. This higher frequency of updates is driving many organizations to rethink how they manage their PC refreshes. In fact, the latest build that Microsoft released is relatively more graphics intensive than previous versions. This is evident from the fact that, in one year alone, Lakeside Software has seen a 20% increase in GPU calls when comparing Windows 10 builds in 2017 to Windows 10 builds in 2018[2].

Figure 1. Comparing Windows 10 Builds in from 2017 to 2018

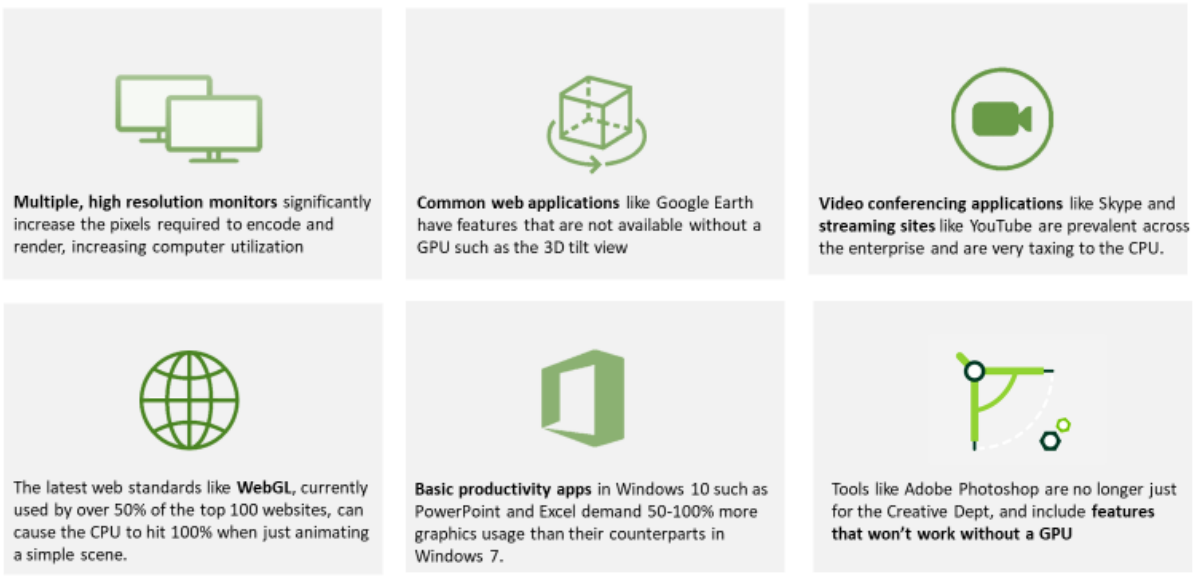
Furthermore, for most of the latest the applications, hardware acceleration is turned on by default. Applications like Office 365, Microsoft teams, PDF viewers and web browsers will generally use hardware acceleration for all WebGL, videos, and zoom, pan, scroll operations.

Desktop virtualization allows IT departments to more easily manage and deploy these new upgrades, but new considerations must be made to accommodate the user experience requirements. Graphics will become even more important when Windows 10 is deployed in virtual environments to ensure that users get full functionality and application compatibility. Nearly everything that the modern digital worker does today has become more graphics

[2] Data from Lakeside Software's SysTrack Community, 2018. Windows 10 builds in 2017 is based on data from builds 1709, 1703 & 1607. Windows 10 builds in 2018 is based on data from builds 1803 and 1709.

intensive, which makes GPUs even more relevant today, as GPUs help offload tasks from the CPU to ensure users get a great experience and can stay productive.

Figure 2. NVIDIA GRID Drives Digital Workplace Transformation



Sizing Guidance for Digital Workplace: Graphics-Intensive Productivity Applications and Windows 10

As you move from a Windows 7 or 8 environment it's important to reevaluate the requirements of your virtual desktops. The frame buffer, or graphics memory, is a dedicated resource in NVIDIA GRID deployments. Understanding the user's workload and related frame buffer requirement can help you determine the overall density of your environment.

Digital worker (a.k.a. knowledge worker) workloads will vary per user depending on many factors, including:

- Number of applications
- The types of applications
- File sizes
- Number of monitors and their resolution

vGPU profiles determine the amount of frame buffer that can be allocated to your virtual machine. For the NVIDIA® M10 GPU with NVIDIA GRID software, the vGPU profiles that are supported are the M10-1B (with 1024 Mbytes of frame buffer) and M10-2B (with 2048 Mbytes of frame buffer). The vGPU profiles of the M10 can support up to four HD, four 2K, or two 4K resolution monitor configurations. Depending on your workload, which manifests vGPU memory requirements, size your vGPU profile correctly by considering all the above factors.

To understand the impact of some of these factors on frame buffer usage, the NVIDIA GRID Performance Engineering team conducted a number of tests using Windows 10 with the NVIDIA M10 GPU.

Based on these tests, NVIDIA recommends that users that have any of the following characteristics should be assigned a GRID Virtual PC (GRID vPC) 1 GB profile to deliver optimal experience:

- Heavy application use - browsing, using email, and creating complex documents, presentations, and spreadsheets.
- Applications used include Windows 10, Office productivity apps, streaming video and multi-media using the latest web standards like WebGL

- Up to two HD (1920 × 1080) or one 2K (2560 × 1440) resolution monitor.

On the other hand, users with any of the following characteristics should be assigned an NVIDIA GRID vPC 2GB profile:

- Users with specific requirements such as multiple, high resolution monitors to handle larger model files and higher resolution media.
- Applications include Windows 10, Office productivity apps, video and multi-media, and industry specific apps like Bloomberg, Thomson Reuters Eikon, and DICOM viewers
- Up to two 2K (2560 × 1440) or two 4K (4096 × 2160) resolution monitors

As with any workload, your results will vary. Therefore, NVIDIA recommends that you conduct real user testing to get specific scalability numbers for your environment.

NVIDIA GRID vPC solution does not target task workers. Task worker workloads are defined as performing repetitive tasks within a small set of apps, usually at a stationary computer. They use non-CPU- and memory-intensive applications and would not benefit from GPUs.

Choosing a vGPU Profile

To deliver optimal performance to your users, it's important to look at several factors to determine the correct vGPU profile. The guidance in the following sections is based on testing by the NVIDIA GRID Performance Engineering team of different factors and their effect on frame buffer usage. The suggested steps for testing within your own environment are provided to help you get the most accurate results.

Frame Buffer Usage per Virtual Machine

Testing was conducted on two different VMs, both of which were configured with 4 vCPUs, 4 GB of vRAM, and NVIDIA Virtual GPU software 7.1 release using H.264 Blast protocol, on Windows 10 Build 1803. The first VM had a GRID vPC license and 1 GB of frame buffer on a NVIDIA M10 (M10-1B profile). The second VM also had a GRID vPC license but the frame buffer was increased to 2 GB on the NVIDIA M10 (M10-2B profile).

Table 1. Frame Buffer Usage

Workload	Resolution	Monitors	Peak FB usage	Median FB usage	Average FB usage	Variance FB usage
M10-1B Profile						
Windows 10 idle	1920x1080 (full HD)	1	245.76MB (24%)			
Knowledge Worker Apps/Win10 active*	1920x1080 (full HD)	1	634.88MB (62%)	430.08MB (42%)	406.29MB (39.67%)	81.25MB (7.93%)
Windows 10 idle	1920x1080 (full HD)	2	409.6 MB (40%)			
Knowledge Worker Apps/Win10 active*	1920x1080 (full HD)	2	839.68 MB (82%)	583.68MB (57%)	563.44MB (55.02%)	80.71 MB (7.88%)
Windows 10 idle	2560×1440 (2K)	1	327.68 MB (32%)			

Workload	Resolution	Monitors	Peak FB usage	Median FB usage	Average FB usage	Variance FB usage
Knowledge Worker Apps/Win10 active	2560×1440 (2K)	1	911.36 MB (89%)	583.68MB (57%)	578.19MB (56.46%)	126.34MB (12.33%)
Windows 10 idle	2560×1440 (2K)	2	512 MB (50%)			
Knowledge Worker Apps/Win10 active	2560×1440 (2K)	2	962.56 MB (94%)	747.52MB (73%)	737.07MB (71.98%)	107.34MB (10.48%)
M10-2B Profile						
Windows 10 idle	4096*2160 (4K)	1	552.96 MB (27%)			
Knowledge Worker Apps/Win10 active	4096*2160 (4K)	1	1679.36 MB (82%)	1146.88 MB (56%)	1059.62MB (51.73%)	252.14MB (12%)
Windows 10 idle	4096*2160 (4K)	2	901.12 MB (44%)			
Knowledge Worker Apps/Win10 active	4096*2160 (4K)	2	1925.12 MB (94%)	1351.68MB (66%)	1328.4 MB (64.86%)	231.15MB (11%)
Windows 10 idle	2560×1440 (2K)	2	614.4 MB (30%)			
Knowledge Worker Apps/Win10 active	2560×1440 (2K)	2	1208.32 MB (59%)	880.64MB (43%)	849.95MB (41.5%)	115 MB (5.61%)

Windows 10 active workload includes a set of applications that are a good representation what knowledge workers widely use. This includes Microsoft Word 2016, Microsoft Excel 2016, Microsoft PowerPoint 2016, and Google Chrome for web browsing, video playback and PDF viewing. The simulated workload also takes into account how these applications are used and is built with human speed input and scrolling. For more information on the workload, refer to our whitepaper on quantifying the impact of virtual GPUs in your VDI environment.

Application Workload

A user's application workload will always be the most important factor in determining which profile will deliver optimal performance. Most users will require 1 GB of frame buffer if they are using applications like Microsoft Office, streaming video and multi-media on Windows 10. 2 GB profile is typical for users with specific requirements such as multiple, high resolution monitors to handle larger files and higher resolution media. These users could also be using Office productivity apps, video and multi-media, as well as industry specific apps like Bloomberg, Thomson Reuters Eikon, and DICOM viewers.

To understand if users can use the GRID vPC with 1 GB of frame buffer (M10- 1B profile), it's important to monitor frame buffer usage with a monitoring tool that exposes GPU resources on their physical PCs or on their VMs. If usage exceeds 90% of 1GB of framebuffer for some

time and average framebuffer utilization exceeds 70%, you should move to 2 GB of frame buffer.

A number of existing monitoring vendors have GPU metrics already built into their platforms. If you do not currently use a monitoring solution, you can use free solutions such as Microsoft Performance Monitor (PerfMon), NVIDIA System Management Interface (nvidia-smi), GPUProfiler, and others.

Monitor Resolution

As monitor resolutions continue to increase, more pixels are being delivered to the screen. As a result, the frame buffer usage in a virtual environment increases.

While full HD is currently the most common resolution, an increasing number of devices are being released with higher resolution screens. On a single 2560 × 1440 monitor, Windows 10 uses 328 MB of frame buffer when idle, about 80 MB more than on a single 1920 × 1080 monitor. When looking at active workloads on a single 2560 × 1440 monitor, Windows 10 uses 911 MB of frame buffer, about 276 MB more than on a single 1920 ×1 080 monitor.

Number of Monitors

The number of monitors to which the VM is being delivered also affects frame buffer usage. Today it is very common to see knowledge workers with at least 2 monitors, which will increase the frame buffer requirements. Likewise, when you increase your monitor count from one to two, Horizon and Citrix remote connections need to encode and render double the number of pixels, thereby increasing CPU utilization.

Testing showed that two full HD monitors running an idle Windows 10 VM used 409 MB of frame buffer. When a workload was added on the VM running two full HD monitors, frame buffer usage reached a peak of 840 MB, which is around 80% of maximum frame buffer utilization of an M10-1B profile. While this gets close to 1 GB of allocated frame buffer, the median is only at 584 MB (57%) and the average is 563 MB (55%) with very little variance or standard deviation seen at 81 MB (8%). Thus, the M10-1B profile is recommended in this scenario to deliver optimal performance.

Testing showed that two 2K (2560 × 1440) resolution monitors running an active workload reached a peak of 963 MB (94%) with median of 748 MB (73%) and average of 737 MB (72%) utilization of an M10-1B profile. This makes using an M10-2B profile more suited for two 2K resolution configurations as this provides more than enough headroom at 1208 MB (59%) peak usage with 880 MB (43%) median and 850 MB (42%) average utilization when running active workloads.

When running active workloads, a single 4K (4096 × 2160) resolution monitor with M10-2B, peak frame buffer utilization is at 1679 MB (82%). While this comes close to the maximum 2

GB allocation, the median is only at 1147 MB (66%) and average at 1060 MB (52%) with low variance of 252 MB (12%) utilization. This makes M10-2B ideal for running single 4K resolution configurations.

When running active workloads on dual 4K (4096 × 2160) resolution monitor with M10-2B, peak frame buffer utilization is at 1925 MB (94%). This likewise comes close to the maximum 2 GB allocation but taking a look at the average frame buffer utilization of 1328 MB and adding the variance of 231 MB gets you to 1559 MB or 76% of frame buffer utilization, which can be acceptable for this configuration.

The 3840 × 1600 or curved screen monitor is a configuration that is becoming more popular as well. This monitor resolution lies in between the 2K and the 4K and as such, we recommend the M10-2B profile for optimal performance.

Note: To provide enough frame buffer to support the application workload, NVIDIA recommends using GRID vPC with 1 GB of frame buffer (M10-1B profile) when using one to two HD (1920 x 1080) or one 2K (2560 × 1440) resolution monitors for knowledge worker workloads. To deliver higher resolutions, two 4K (4096 × 2160) or two 2K (2560 x 1440), you should use 2 GB of frame buffer (M10-2B profile) for the best performance.

Managing at Scale

When deploying VDI across your enterprise, it is important to have the right tools to manage your environment at scale. XenMotion for live migration of GPU accelerated VMs is now supported with the NVIDIA virtual GPU 6.0 release, and VMware vMotion technology is now supported with the NVIDIA virtual GPU 7.0 release.

IT admins can now use vMotion or XenMotion to live migrate NVIDIA vGPU-powered virtual desktops without end-user interruption and with no data loss.

This allows IT to perform critical services like workload leveling, infrastructure resilience and server software upgrades without any VM downtime, when it's convenient for them. With live migration, IT can truly deliver high availability and a quality user experience.

Using a 1 GB profile, migration of a VM from one host to another just takes around a few seconds, and just 2 clicks.

Testing Your NVIDIA GRID Environment

As with all scalability testing, user workloads and environment must be taken into account. In order to test NVIDIA GRID in your environment you can choose to get started with a certified NVIDIA partner or start a proof of concept (POC) with a certified server and the 90-day NVIDIA GRID evaluation license.

The following are important things to remember during your POC.

1. Define "acceptable" user experience.
   Defining user experience (UX) requires careful examination of user and application interaction.
   A definition of acceptable user experience can be obvious, such as the rendering time for an image to appear, or the ability to smoothly pan across that image. It can also be less obvious, such as the ability to smoothly scroll down a page or the "snappy" reaction for a menu to appear after a right click.

Note: Understanding the limitations of synthetic benchmarks.

Understanding the trade-off between the level of user experience and resource utilization will improve the accuracy in design and sizing decisions for the VDI infrastructure. However, existing benchmarks today do not fully capture user experience as they do not focus on the client or receiver side from where an end user interacts with a remote session.

To address the need for better insights of the actual end user experience, NVIDIA developed a benchmarking tool based on a methodology that measures key aspects of the user experience, including end user latency, consistency of user experience, remoted frames, image quality, and server utilization. The results of our testing showed a GPU-enabled VDI environment significantly improves the user experience as well as the server density when compared to a CPU-only VDI environment. Some of the benefits include up to 3X improved end user latency, up to 1.5X more remoted frames, and 33% more users supported on the server. For more information, refer to our whitepaper.

2. Compare real-world workloads.
   The more your workload is closer to what actual users do, the more accurate your sizing decisions will be. Aspects like remote session interactivity, color accuracy, having multiple screens, and ensuring realistic emulation of end-user behavior are often not captured in legacy benchmarks.
   To see how NVIDIA's benchmarking approach differs from legacy benchmarking tools, refer to Appendix A.

3. Test with real users.
   It's important to look at the application running to be sure that the experience is enjoyable for users.
   While idle systems or benchmarks can be used as a starting point, real user workflows may require more or less graphics resources. Because the number of monitors and their resolution will impact graphics requirements, users bringing their own devices or logging in from different locations may impact the overall user experience.
   This is why it is important to conduct your own side-by-side testing. There is nothing like seeing the difference for yourself, in a test environment that is representative of what your users do on a day-to-day basis. To avoid generic feedback, ask users to report metrics, and to judge specific activities or functions using finite scales (for example, 1-5, 5 being best). For more information on how to setup your own test environment, refer to our testplan.

Appendix A

Behind the Numbers: Ten Variables Evaluated by NVIDIA's Benchmarking Approach

For any VDI deployment, the goal is to always enable end users with a local desktop-like user experience while bringing the full spectrum of proven benefits of VDI. Moving away from a physical PC or workstation to a VDI enabled desktop or remote session should not compromise performance and productivity.

There is a need to objectively measure the user experience of a VDI desktop or remote session. Understanding the trade-off between the level of user experience and resource utilization improves the accuracy in design and sizing decisions for the VDI infrastructure.

The recommended best practice when evaluating user experience is to focus on the client or receiver side from where an end user interacts with a remote session. Following, we enumerate NVIDIA's ten areas of focus while quantifying user experience in context with resource utilization.

1. Remote session interactivity as measured by end user latency
   End user latency is an integral part of quantifying user experience. It measures the time taken to observe the response once an input is sent to the application running in the VDI session.
2. Color accuracy as measured by image quality using SSIM algorithm
   Color accuracy, brightness, sharpness of foreground and background colors directly influence user experience. We measure the fidelity between screens rendered by the VDI instance and the screens observed by the remote user. Calculating 'structural similarity' quantifies color accuracy.
3. Session consistency as measured by remoted frames
   Consistency in delivering a good user experience plays a big role. The number of frames a remoting protocol sends over to the end user is a driver of how fluid and non-jittery a remote session feels. Measuring the encoded frames by the remoting protocol gives us a metric in determining smoothness of a remote session.
4. Multiple screens, multiple screen resolutions
   Use of multiple, high resolution monitors such as 1080p and 4K are now becoming

mainstream. Evaluating user experience with multiple monitors and higher screen resolutions provides a more accurate representation of how people work today.

5. One virtual client per virtual desktop
   One remote session per virtual client to a virtual desktop is more realistic of real-world deployments. This is the only way to measure user interactivity and color accuracy which are central to quantifying user experience.
6. Realistic emulation of end-user behavior
   A workload emulating the most common activities that a real user does like human speed scrolling, zooming and multi-application interactivity gives a better sense of what the user experience will be once your VDI deployment goes live. A realistic modeling of the behavior of a typical VDI user leads to more accurate sizing recommendations.
7. Universal framework to evaluate performance and user experience
   Some applications use a variety of graphics APIs such as OpenGL, variants of DirectX and modern web standards like WebGL, while some applications do not. A single framework that accounts for a broad range of applications is needed. The impact of GPUs needs to be quantified with a workload that factors a breadth of applications used in the modern digital workplace.
8. Measuring performance stats on the platform
   Measuring performance stats on the hypervisor, GPU hardware and the guest OS that is running VDI enables user experience to be quantified in context of resource utilization. Understanding this trade-off enables higher accuracy in sizing and design decisions while architecting a VDI infrastructure.
9. Provisioning virtual clients along with virtual desktops
   The framework taking over some of the burden in provisioning the test environment enables focus on evaluating user experience and its trade-off with resource utilization. Automated provisioning and management of virtual machines needed for VDI desktops and virtual machines that act as end-points enable huge time savings. This avoids human error in creating an environment for performance evaluation.
10. Specific workloads for a specific use case
    Designers and engineers use high end 3D applications, such as Autodesk AutoCAD, Siemens NX or Dassault Systèmes SOLIDWORKS, etc. Knowledge workers, on the other hand, use modern productivity applications like web browsers, Microsoft Office application suite or Adobe® Acrobat®. Having workloads for specific use cases lets us more precisely evaluate user experience and platform performance.

Improve the VDI Experience of Your Users

It's important to properly evaluate user experience as IT designs their infrastructure and thinks about broad scale deployment. By using a combination of metrics from NVIDIA's benchmarking tool along with detailed user feedback, IT administrators can better quantify VDI user experience.

Notice

The information provided in this specification is believed to be accurate and reliable as of the date provided. However, NVIDIA Corporation ("NVIDIA") does not give any representations or warranties, expressed or implied, as to the accuracy or completeness of such information. NVIDIA shall have no liability for the consequences or use of such information or for any infringement of patents or other rights of third parties that may result from its use. This publication supersedes and replaces all other specifications for the product that may have been previously supplied.

NVIDIA reserves the right to make corrections, modifications, enhancements, improvements, and other changes to this specification, at any time and/or to discontinue any product or service without notice. Customer should obtain the latest relevant specification before placing orders and should verify that such information is current and complete.

NVIDIA products are sold subject to the NVIDIA standard terms and conditions of sale supplied at the time of order acknowledgement, unless otherwise agreed in an individual sales agreement signed by authorized representatives of NVIDIA and customer. NVIDIA hereby expressly objects to applying any customer general terms and conditions with regards to the purchase of the NVIDIA product referenced in this specification.

NVIDIA products are not designed, authorized or warranted to be suitable for use in medical, military, aircraft, space or life support equipment, nor in applications where failure or malfunction of the NVIDIA product can reasonably be expected to result in personal injury, death or property or environmental damage. NVIDIA accepts no liability for inclusion and/or use of NVIDIA products in such equipment or applications and therefore such inclusion and/or use is at customer's own risk.

NVIDIA makes no representation or warranty that products based on these specifications will be suitable for any specified use without further testing or modification. Testing of all parameters of each product is not necessarily performed by NVIDIA. It is customer's sole responsibility to ensure the product is suitable and fit for the application planned by customer and to do the necessary testing for the application in order to avoid a default of the application or the product. Weaknesses in customer's product designs may affect the quality and reliability of the NVIDIA product and may result in additional or different conditions and/or requirements beyond those contained in this specification. NVIDIA does not accept any liability related to any default, damage, costs or problem which may be based on or attributable to: (i) the use of the NVIDIA product in any manner that is contrary to this specification, or (ii) customer product designs.

No license, either expressed or implied, is granted under any NVIDIA patent right, copyright, or other NVIDIA intellectual property right under this specification. Information published by NVIDIA regarding third-party products or services does not constitute a license from NVIDIA to use such products or services or a warranty or endorsement thereof. Use of such information may require a license from a third party under the patents or other intellectual property rights of the third party, or a license from NVIDIA under the patents or other intellectual property rights of NVIDIA. Reproduction of information in this specification is permissible only if reproduction is approved by NVIDIA in writing, is reproduced without alteration, and is accompanied by all associated conditions, limitations, and notices.

ALL NVIDIA DESIGN SPECIFICATIONS, REFERENCE BOARDS, FILES, DRAWINGS, DIAGNOSTICS, LISTS, AND OTHER DOCUMENTS (TOGETHER AND SEPARATELY, "MATERIALS") ARE BEING PROVIDED "AS IS." NVIDIA MAKES NO WARRANTIES, EXPRESSED, IMPLIED, STATUTORY, OR OTHERWISE WITH RESPECT TO THE MATERIALS, AND EXPRESSLY DISCLAIMS ALL IMPLIED WARRANTIES OF NONINFRINGEMENT, MERCHANTABILITY, AND FITNESS FOR A PARTICULAR PURPOSE. Notwithstanding any damages that customer might incur for any reason whatsoever, NVIDIA's aggregate and cumulative liability towards customer for the products described herein shall be limited in accordance with the NVIDIA terms and conditions of sale for the product.

Adobe

Adobe, the Adobe logo, and Acrobat are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States and/or other countries.

Trademarks

NVIDIA, the NVIDIA logo, and NVIDIA GRID are trademarks and/or registered trademarks of NVIDIA Corporation in the U.S. and other countries. Other company and product names may be trademarks of the respective companies with which they are associated.

Copyright

NVIDIA Corporation | 2788 San Tomas Expressway, Santa Clara, CA 95051
http://www.nvidia.com