Application Deployment Guide

Autodesk AutoCAD with NVIDIA GRID™ vGPU™ on VMware Horizon

AUTODESK. | NVIDIA. | vmware

THE QUESTION

HOW MANY USERS CAN I GET ON A SERVER?

This is a typical conversation we have with customers considering NVIDIA GRID vGPU:

How many users can I get on a server?

NVIDIA: What is their primary application?

Autodesk AutoCAD 2015.

NVIDIA: Are they primarily architects or designers?

Designers mostly.

NVIDIA: Are they doing mostly 2D or 3D design work?

Mostly 3D.

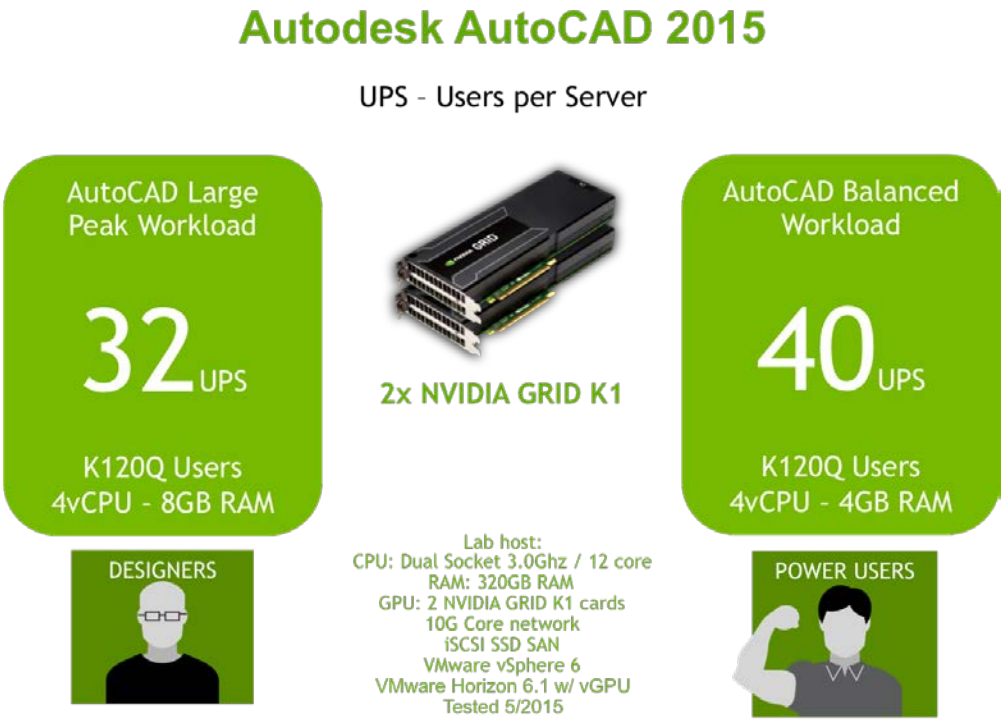
NVIDIA: Power users to designers then.

I need performance AND scalability numbers that I can use to justify the project.

THE ANSWER – USERS PER SERVER (UPS)

UPS – USERS PER SERVER

Based on the NVIDIA GRID™ Performance Engineering Lab (GRID P.E.L.) findings, NVIDIA GRID provides the following performance and scalability metrics for Autodesk AutoCAD 2015. These metrics are based on tests with the lab equipment shown in the graphic below, using the Cadalyst benchmark, and in working with Autodesk and their emphasis on end user usability. Of course, your usage will depend on your models and equipment, so this Application Guide is intended to be used as a starting point for your implementation.



ABOUT THE APPLICATION: AUTOCAD 2015

Autodesk AutoCAD is one of the most globally used software platforms for CAD design and documentation. AutoCAD leverages both CPU and GPU to deliver a high quality user experience, and as a result, there are several things that need to be considered in architecting your GRID vGPU solution: the size of your map data, the concurrency of your users, and the level of interaction with 3D data.

User Classification Matrix

AutoCAD's user classification is shown in Table-01, along with a mapping to the NVIDIA user classifications as reference:

Table-01

User Classification Matrix			
NVIDIA User Classifications	**Knowledge Worker**	**Power User**	**Designer**
AutoCAD User Classifications	Basic/Mobile	Architect	Power User
AutoCAD Build Spec	Minimum	Recommended	3D & Large Data Sets

HOW TO DETERMINE USERS PER SERVER

This section contains an overview of the GRID P.E.L., recommended virtual desktop builds, the testing methodology used, and the metrics and results that support the findings in this application guide.

THE NVIDIA GRID PERFORMANCE ENGINEERING LAB

The GRID P.E.L.'s mandate is to measure and validate the performance and scalability delivered via the GRID platform (GRID vGPU software running on NVIDIA GRID GPUs) on all enterprise virtualization platforms. It is our goal to provide proven testing that gives customers the ability to build and test their own successful deployments.

The GRID P.E.L. works closely with our counterparts in the enterprise virtualization community, including ISVs, OEMs, vendors, partners, and their user communities to

determine the best methods of benchmarking in ways that are both accurate and reproducible.

Leveraging our lab for enterprise virtualization technology, the GRID P.E.L. has the capacity to run a wide variety of tests ranging from standard benchmarks to reproducing customer scenarios across a wide range of hardware, including different OEM servers with varying CPU specifications, storage options, client devices, and network configurations.

The goal is to find the most accurate proxy possible for testing; however, this is still not the same as real users doing real work with their data. The GRID P.E.L. is committed to working with customers to find more and better models, and field confirmation of findings.

TYPICAL AUTOCAD WORKSTATION BUILDS

Autodesk delivers a recommended hardware specification to help choose a physical workstation. These recommendations provide a good starting point for your virtual desktops. Based on our Cadalyst testing results, along with feedback from early customers, these are our recommended virtual system requirements. Your own tests with your own models will determine if these recommendations meet your specific needs.

Recommended AutoCAD Virtual System Requirements

Working with VMware and our shared customers with their tested or production environments, we are recommending the following system requirements for deploying AutoCAD in a virtual environment:

VMware: Recommended Configuration			
VMware Software	VMware vSphere 6 or later VMware Horizon 6.1 or later		
Virtual Machine Operating System	Microsoft Windows 7 SP1 64-bit: Enterprise, Ultimate, or Professional Microsoft Windows 8.1 64-bit: Enterprise, or Professional		
Host Server	Minimum	Recommended	3D & Large

Recommendation			Datasets
CPU	2.6 GHz+ Intel Xeon E5 or later	3.0 GHz+ Intel Xeon E5 or later	3.0 GHz+ Intel Xeon E5 or later
Memory	196 GB	320 GB	320 GB
Networking	1 GB minimum 10 G recommended	10 G	10 G
Storage	~250+ IOPS Per User	~500+ IOPS Per User	~750+ IOPS Per User
GPU	NVIDIA GRID K1 or greater NVIDIA GRID K2 or greater	NVIDIA GRID K1 or greater NVIDIA GRID K2 or greater	NVIDIA GRID K1 or greater NVIDIA GRID K2 or greater
Virtual Machine Settings	Minimum	Recommended	3D & Large Datasets
Memory	4 GB RAM	6 GB RAM	8 GB RAM or greater
vCPUs	2+ vCPUs	4 vCPUs	4 vCPUs
Disk Space	25 GB free disk space	50 GB free disk space	100 GB free disk space
Graphics Adapter	NVIDIA GRID K120Q (512 MB) or greater NVIDIA GRID K220Q (512 MB) or greater	NVIDIA GRID K120Q (512 MB) or greater NVIDIA GRID K220Q (512 MB) or greater	NVIDIA GRID K140Q (1 GB) or greater NVIDIA GRID K240Q (1 GB) or greater

For these tests, the GRID P.E.L. keys on recommended specifications when feasible. The goal is to test both performance and scalability while maintaining the flexibility and manageability advantages of virtualization without sacrificing the performance end users expect from NVIDIA-powered graphics.

UX – THE VDI USER EXPERIENCE

Defining user experience (UX) requires careful examination of user and application interaction. This can be obvious, like the rendering time for an image to appear, or smoothly panning across that image. It can also be less obvious, like the ability to smoothly scroll down a page or the "snappy" reaction for a menu to appear after a right click. While elements such as these can be measured, the user's perception is much harder to measure.

Users also add variables like "think time", the time they spend looking at their display before interacting again with the application. This time offers an advantage to the underlying resources, such as CPU, as it allows tasks to finish and processes to complete. It is even more beneficial in a shared resource environment such as VDI, where one user "thinking" frees up resources for another user who chose that moment to interact with their application. Now factor in other time away from the application (meetings, lunch, etc.) and one could expect to get even more benefits from shared resources. These benefits equate to more resources for the user's session and typically a more responsive application, thus a better-perceived experience by the end user.

As user experience is typically defined in terms of what an end user receives as a visual experience, GRID P.E.L. used frames per second (FPS) to measure results. By using FPS, we can analyze when a user would "see" a reduction in performance as load increases on the system. The larger AutoCAD community uses a benchmark called Cadalyst, which is explained in detail in the next section, and see the following document for results using the Cadalyst Index:

http://www.nvidia.com/content/grid/resources/AutoCAD_GRID_vGPU_Scalability_Solutions_Guide.pdf

CADALYST Benchmark Metrics

The extended AutoCAD community has generally relied on a benchmark called Cadalyst. It interacts with the application and an accompanying model to run several tests, then compares those results against a known baseline, and reports the results. The benchmark is available here: http://www.cadalyst.com/benchmark-test

From the Cadalyst website: The Cadalyst Systems Benchmark is designed for use with full, basic AutoCAD only. It is not intended for use (and may not function properly) with market-specific AutoCAD versions such as Architecture, Mechanical, or Civil 3D; AutoCAD LT; trial versions; international versions; etc.

The test evaluates four areas of system performance: 3D Graphics, 2D Graphics, Disk,

and CPU. It compares the test times of your current system with a set of base times and computes a Total Index Score. As is true for all benchmark tests, the Cadalyst Benchmark Test is merely a guide for measuring system performance.

The 3D visualization portions of the benchmark are compatible with AutoCAD 2009 and later; the 2D portions of the test work with AutoCAD 2000 and later. An index score of 135, for example, means your test system is 135 times faster than the base system for the functions tested.

Real-Life User Experience Metrics

It's important to actually see the tests being run to be sure that the experience is enjoyable for users. That being said, it's also important to maintain perspective, especially if you are not a regular user of applications like AutoCAD. While a data center admin deploying an AutoCAD in a VDI workload might view a testing desktop and think the experience is slow or sluggish, a user who works in it daily might find it normal. The feedback from an actual 3D designer using the application in a virtual desktop is the ultimate test of success.

TESTING METHODOLOGY

To ensure that test results were repeatable, we deliberately chose the Cadalyst Benchmark workload and executed simultaneous tests, meaning all testing virtual desktops were executing the same activities at the same time. This "peak workload" scenario is a worst-case representation of real user interaction, where the results show the number of users per host when the highest possible load generated by the application is put on the shared resources.

- Sample workload: Cadalyst provides their workload, a set of models, for us to test with.
- Scripting: As Cadalyst is historically designed for single physical workstation testing, there is no built-in automation for multi-desktop scalability testing. As a result we use tools like Login VSI, AutoIT, and similar.
- Think Time: By adding a length of time between tests, we are making a basic effort to create synthetic human behavior.
- Staggered Start: By adding a delay to the beginning of each test, we are offsetting the impact of tests were they to be run in unison, again an effort to create synthetic human behavior.

- Scalability: In general, we ran 1 virtual desktop, then 8, then 16, and so on, to get a baseline of results and accompanying logs (CPU, GPU, RAM, networking, storage IOPS, etc.).

RESULTS

The following are results of our testing, looking for the greatest scalability while still within performance expectations. It's important to note that your users, your data, and your hardware will impact these results, and you may decide a different level of performance or scalability is required to meet your individual business needs.

As the Catalyst benchmark does not push AutoCAD's GPU capabilities, and was built to push the limits of dedicated hardware versus the shared resources of VDI, the decision was made to stop testing once the CPU was approaching 100% utilized and test times had climbed past twice what we were finding on the a single physical workstation with dedicated resources. We then met with the Autodesk AutoCAD team, discussed the results, reviewed the tests in action, and physically verified that this was still within what a typical user would deem acceptable and usable.

When running AutoCAD, anything higher than 20 FPS is awesome, but users generally don't notice the difference once you exceed 30 FPS. However, once you drop below 10 FPS, the software is going to feel very sluggish and become unusable by the time you hit 5 FPS.

- 20 fps above is good
- Below 10fps – sluggish
- 5 fps – unusable

32VM : K220Q perform better than K120Q but K120Q is still maintain avg 31FPS.
40VM : Due to vGPU configuration, need K120Q, avg FPS is 24 FPS
48VM : one you hit 40 VM running, avg FPS is dropping below 10 FPS.

It's been well documented that storage performance is key to providing high performance graphics workloads, especially with many users and ever-growing file or model sizes. In our lab. we were using a 10G iSCSI connected all flash SAN from Pure Storage. At no time in these tests were IOPS an issue, but it's important to note that as you scale to multiple servers hosting many guests, this needs to be monitored.

Below are our results with analysis, first on Intel Ivy Bridge processors, then on Haswell. Lower scores are better, representing less time to perform the activity.

Table 1: K220Q performs better than K240Q for Haswell servers (for the Cadalyst models).

Table 2: Significant improvements using Haswell over Ivy bridge – Although Ivy bridge clock speed is higher (3.0 GHz to 2.3 GHz), Haswell also has higher number of logical processors (64 vCPUS) when compared to IB (40), but there is improvement for under-utilized host as well.

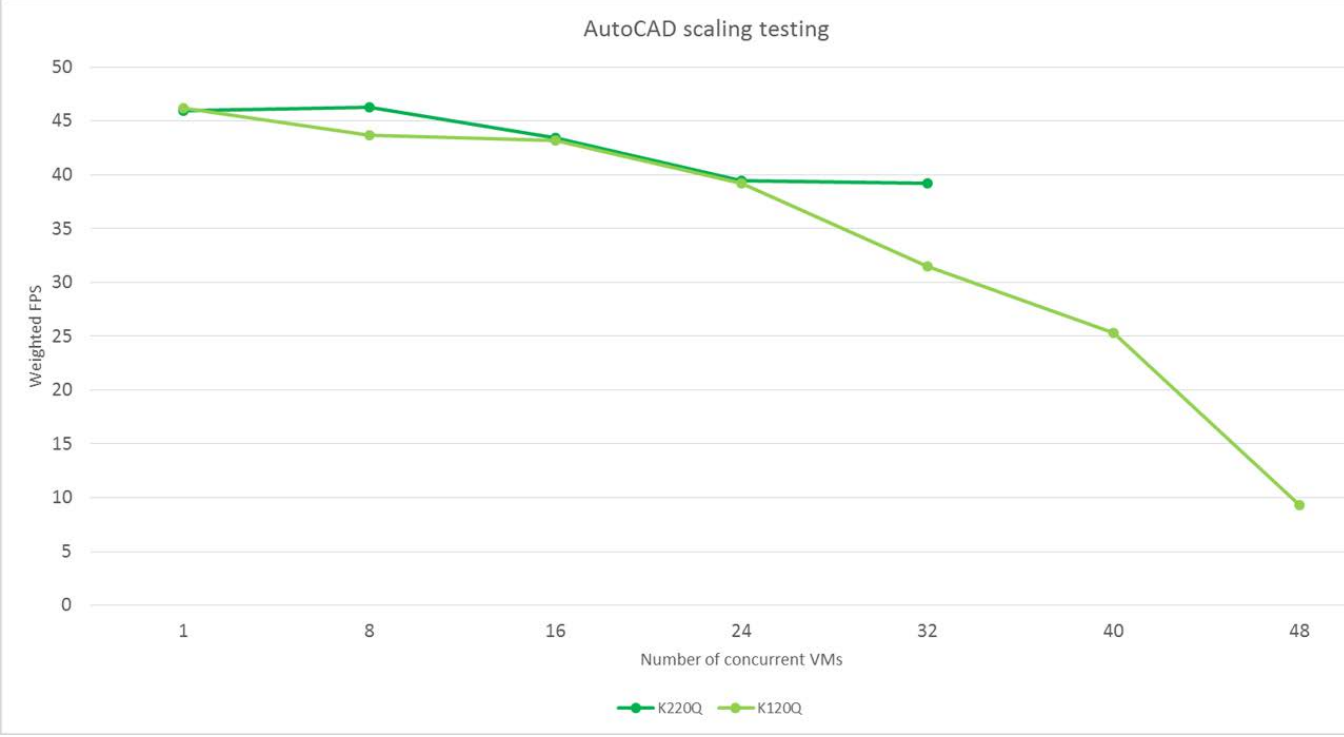
Tables 3 & 4: Due to the staggered start, numbers are slightly better than the previous test runs.

For all cases, CPU only tests (without hardware acceleration) for a single VM takes more time than 16/24 VM tests with GPU (with hardware acceleration). This could be due to virtualization overhead. Since these models are quite small, managing larger GPU memory might have a penalty associated with it.

Results Summary

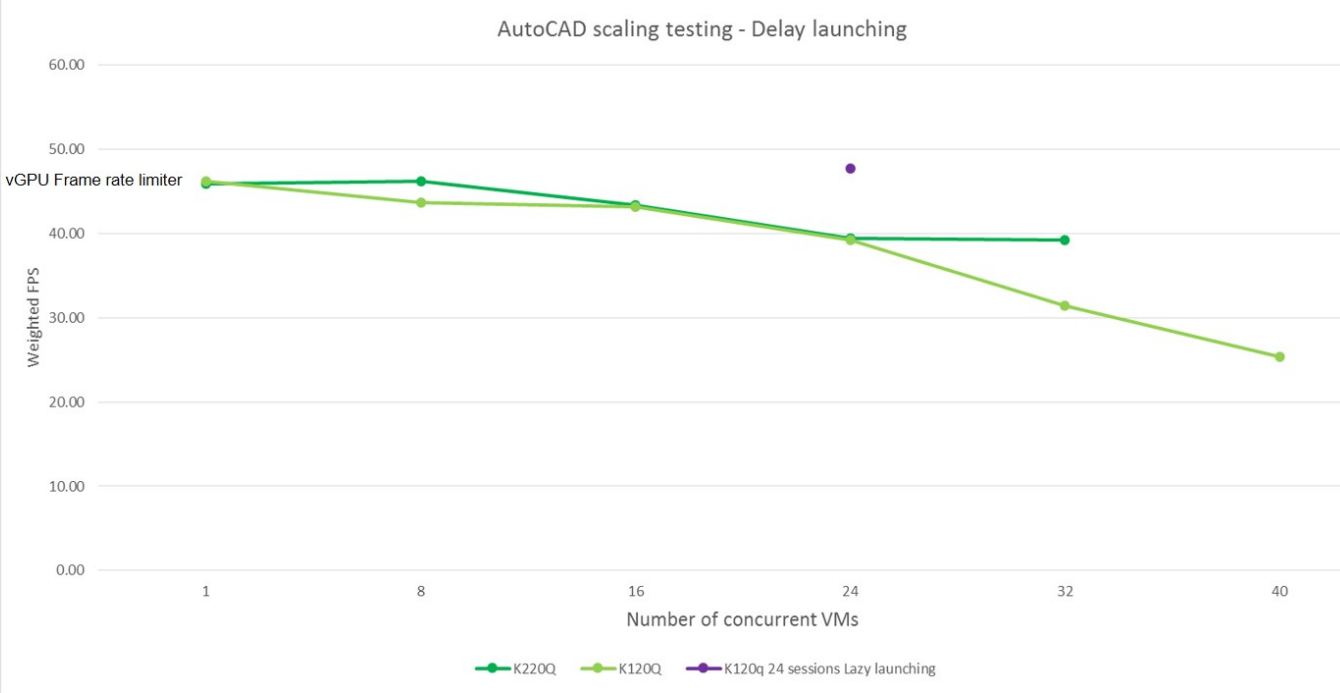
AutoCAD requires significant CPU resources, so investing in higher CPU speeds and more cores will pay off on performance and scalability. For medium to large models, K240Q performance might be better for a real use scenario. Since your own situation is different, you must test with your own models to ensure the most accurate results.

The chart below shows a comparison between two different sets of VDI guests, one using the K2 and its K220Q profile, the other a K1 and its similarly sized frame buffer profile, the K120Q:



Notice that as scale, the number of concurrent VM's increases, there is an obvious drop in performance, and logs will show this is the CPU becoming 100% utilized. This is a "peak workload" test as all sessions are started in unison, we would expect humans to be more staggered in their workflow.

The next chart shows results when we add that synthetic human behavior, we delay the launching of each session by 5 seconds, this offsets the tests and prevents all tests hitting the same function and impacting the CPU in unison.



Notice that with delayed launching, or staggered starts, the performance remains high to a much greater scale of concurrent sessions. Additional findings from this testing:

- vCPU - 2 vCPU are sufficient for this benchmark, however increasing model size and complexity will change this.
- Virtual System Memory – 2 GB is sufficient for this benchmark, however increasing model size and complexity will change this.
- K220Q/K120Q : both show capped FPS caused by frame rate limiting (FRL)
- vGPU has 45 fps frame limiter for performance balancing across multiple VM.

The following table shows the weighted FPS results for various vCPU count and vRAM amounts:

vCPU	RAM	vGPU	Weighted FPS
1	2GB	K220Q	32.27
2	**2GB**	**K220Q**	**45.21**
2	2GB	K100	37.31
2	**2GB**	**K120Q**	**46.17**
2	**2GB**	**K140**	**44.03**
4	1GB	K120Q	48.51
4	2GB	K220Q	47.06
4	4GB	K220Q	45.94
6	4GB	K220Q	46.47
8	6GB	K220Q	46.15

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 regard 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.

VESA DisplayPort

DisplayPort and DisplayPort Compliance Logo, DisplayPort Compliance Logo for Dual-mode Sources, and DisplayPort Compliance Logo for Active Cables are trademarks owned by the Video Electronics Standards Association in the United States and other countries.

HDMI

HDMI, the HDMI logo, and High-Definition Multimedia Interface are trademarks or registered trademarks of HDMI Licensing LLC.

ROVI Compliance Statement

NVIDIA Products that support Rovi Corporation's Revision 7.1.L1 Anti-Copy Process (ACP) encoding technology can only be sold or distributed to buyers with a valid and existing authorization from ROVI to purchase and incorporate the device into buyer's products.

This device is protected by U.S. patent numbers 6,516,132; 5,583,936; 6,836,549; 7,050,698; and 7,492,896 and other intellectual property rights. The use of ROVI Corporation's copy protection technology in the device must be authorized by ROVI Corporation and is intended for home and other limited pay-per-view uses only, unless otherwise authorized in writing by ROVI Corporation. Reverse engineering or disassembly is prohibited.

OpenCL

OpenCL is a trademark of Apple Inc. used under license to the Khronos Group Inc.

Trademarks

NVIDIA, the NVIDIA logo, NVIDIA GRID, and NVIDIA GRID vGPU 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