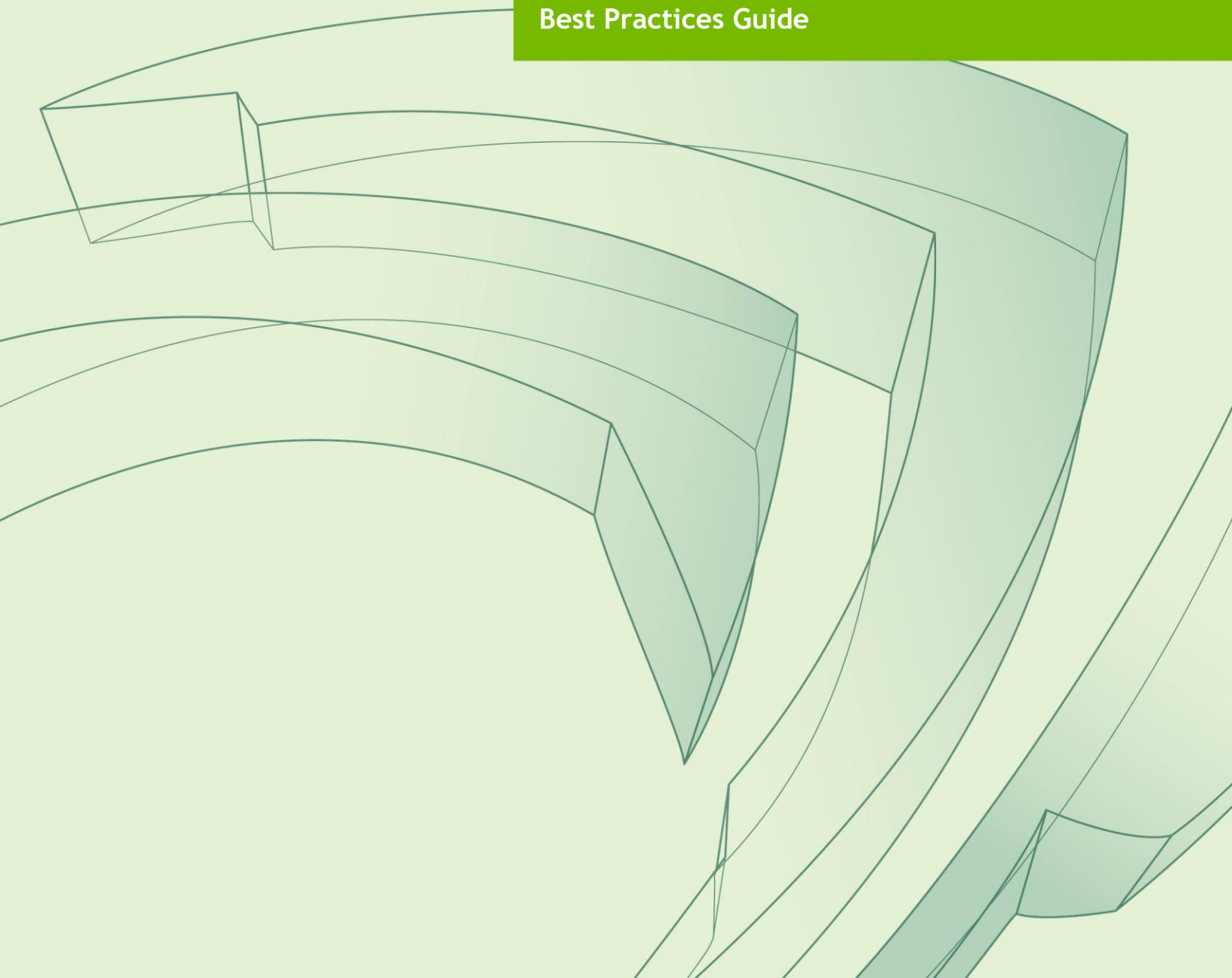
NVIDIA GRID vGPU PROFILE SIZING FOR WINDOWS 10

January 2017
NVIDIA Performance Engineering

Best Practices Guide



HOW TO CHOOSE THE CORRECT NVIDIA GRID vGPU PROFILE FOR WINDOWS 10

NVIDIA GRID vGPU AND MICROSOFT WINDOWS 10

To provide the right level of user experience for your users, 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 guide is to provide guidance on how to assign the appropriate amount of frame buffer for Windows 10 knowledge worker use cases in NVIDIA GRID™ environments.

The latest operating system from Microsoft, Windows 10, was designed to deliver improved user experience on both PCs and mobile devices. 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. 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.

SIZING GUIDANCE FOR WINDOWS 10

Windows 10 is proving to be the most graphics-intensive operating system that Microsoft has ever released. As you move from a Windows 7 or 8 environment it's important to reevaluate the requirements of your virtual desktops as you transition. Frame buffer, or graphics memory, is a dedicated resource in NVIDIA GRID deployments and can help you determine the overall density of your environment.

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

To size your vGPU profile correctly, all of these factors must be considered.

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 Tesla™ M10 GPU. Based on these tests, NVIDIA recommends that users that have any of the following characteristics should be assigned a vPC 1 GB profile to deliver optimal experience:

- Heavy application use; including WebGL, video streaming, and flash applications
- 2560×1600 resolution or higher
- 2 or more 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.

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 VM

Testing was conducted on two different VMs, both of which were configured with 2 vCPUs, 4 GB of vRAM, and NVIDIA GRID 4.1 software. The first VM had a Virtual PC (vPC) license and 512 MB of frame buffer on a Tesla M10 (M10-0B profile).The second VM also had a vPC license but the frame buffer was increased to 1 GB on the Tesla M10 (M10-1B profile).

Workload	Resolution	Monitors	Frame buffer usage
M10-0B Profile			
Windows 7 idle	1920x1080 (full HD)	1	168 MB
Windows 10 idle	1920x1080 (full HD)	1	229 MB
Windows 10 idle	2560x1600	1	287 MB
LoginVSI Win10 Knowledge Worker	1920x1080 (full HD)	1	411 MB
M10-1B Profile			
Windows 10 idle	1920x1080 (full HD)	2	283 MB
LoginVSI Win10 Knowledge Worker	1920x1080 (full HD)	2	664 MB

Application Workload

A user's application workload will always be the most important factor in determining which profile will deliver optimal performance. Office workers today use many applications simultaneously, often with multiple applications interacting with one another. Based on a study with Lakeside Software, the number of applications that use OpenGL or DirectX, and, therefore leverage graphics acceleration, has doubled in the last 5 years. Today, almost 60% of users have at least one application that is graphics accelerated.

To simulate application workloads in the testing, the LoginVSI Windows 10 Knowledge Worker (medium) workload was used. Like any benchmark, this synthetic workload can be used to simulate real user behavior but should not take the place of real user testing.

To understand if users can use the Virtual PC edition with 512 MB of frame buffer (M10-0B 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 512 MB of frame buffer, you should move to 1 GB of frame buffer (M10-1B profile).

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. Today, most users have full HD (1920×1080) resolution or above, which uses a minimum of 229 MB of frame buffer when Windows 10 is idle.

While full HD is currently the most common resolution, an increasing number of devices are being released with higher resolution screens. On a 2560×1600 monitor, Windows 10 uses 287 MB of frame buffer when idle, about 60 MB more than on a 1920×1080 monitor.

To provide enough frame buffer to support the application workload, NVIDIA recommends using Virtual PC with 1 GB of frame buffer (M10-1B profile) when using 2560×1600 resolution for office workloads. To deliver higher resolutions, up to 4K (4096×2160), you should use Virtual Workstation (vWS) with a minimum of 1 GB of frame buffer (M10-1Q profile).

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 office workers with at least 2 monitors, which will increase the frame buffer requirements.

Testing showed that 2 full HD monitors running an idle Windows 10 VM used 283 MB of frame buffer. When a workload was added on the two full HD monitors, frame buffer usage reached 611 MB, which exceeds the 512 MB of the M10-0B profile.

When you are supporting a knowledge worker environment that has two or more monitors, NVIDIA requires that you use Virtual PC with 1 GB of frame buffer (M10-1B profile). Depending on your overall workload, 3-4 monitors may require using Virtual Workstation with 2 GB of frame buffer (M10-2Q) to deliver optimal performance.

TESTING YOUR NVIDIA GRID ENVIRONMENT

As with all scalability testing, user workloads and environment must be taken into account for scalability analysis. 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.

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.

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).

2. **Compare real-world workloads.**

   In virtual environments, time-slicing of resources allows users to get the same level of performance even when sharing resources.

   Time-slicing results from users' thinking time, which includes any pause in their interaction with the application, any period when they are not using the application, or even sitting at their desks.

   By adding up all the time away from the application (for example, meetings, lunch, and periods out of office) you could expect to get even more benefits from shared resources. These benefits equate to more resources for each user's session and typically a more responsive application.

   A realistic estimate of user interaction with an application results in a better-perceived experience by the end user than benchmarks based on peak workloads with inhuman work that is unrepresentative of real users' interactions with an application.

3. **Test with real users**.

   It's important to actually 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.

   When you also consider the effect of real-world workloads, you can see why real users are the most accurate means of testing.

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