Dell Wyse Datacenter

Getting the best value from 3D Graphics in a VDI environment

2/25/2014
Phase 1
Version 1.2

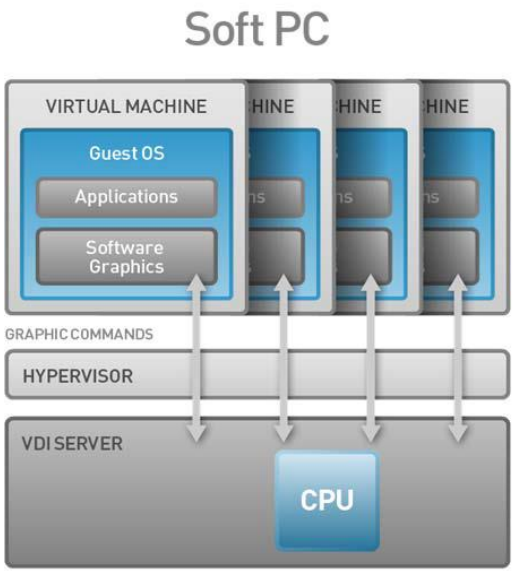
Getting the best value with 3D Graphics

Delivering graphics in a virtualized desktop environment has typically involved one of two approaches-delivering high quality graphics (direct graphics) with such a low user density making the solution very cost-prohibitive. The other option available was low end graphics (shared graphics) with a high enough user density to make the solution economically viable.

Over the next few paragraphs, we will explore the historic graphics options that were available to VDI Admins and a revolutionary new "hybrid" technology (NVIDIA GRID™ vGPU™) that combines the best of both worlds-offering high-end 3D graphics without sacrificing, and in fact improving upon, user density.

The Historical stumbling blocks to GPU-Accelerated VDI

With so many compelling advantages to anywhere/any-device remote visualization, graphics-rich VDI should be more commonplace, but it's not. While VDI has found broad acceptance in mainstream computing, serving graphics-rich, workstation caliber applications and data from a server to a remote client is found today only in relatively small niches. Why? Simply put, first-generation GPU-accelerated solutions either fell short in their performance or disappointed in their versatility. The simplest and most accepted way to implement VDI to date is though a fully abstracted software implementation of a virtual machine, running on the server: the Soft PC. With no GPU present, the CPU has to process all of the workload, including the graphics, a lot like that mainframe/terminal combination of decades ago.

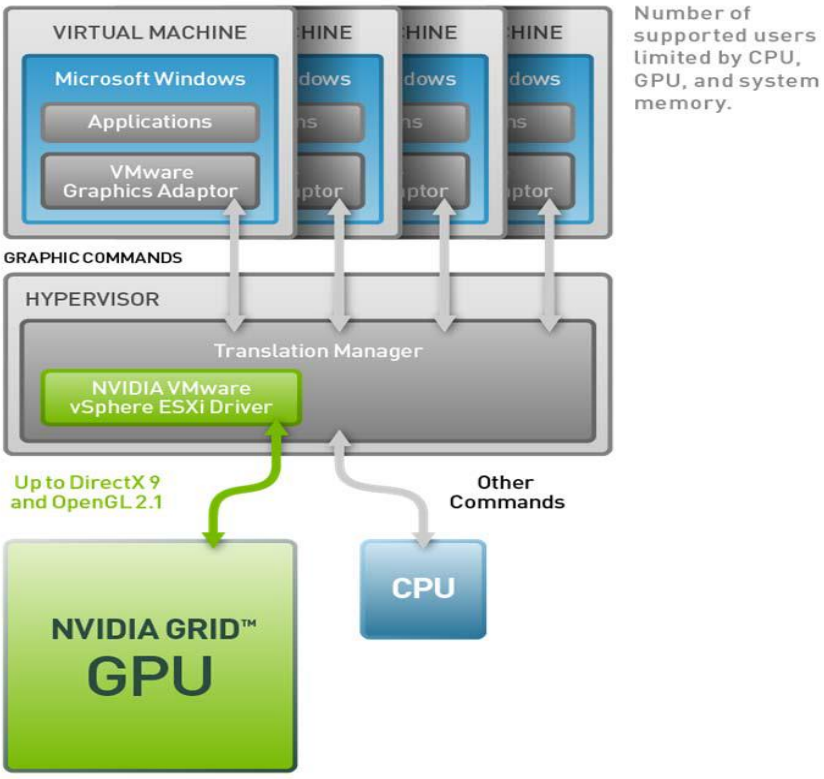


While a Soft PC implementation can work fine for simple text-based, console type applications, it cannot deliver an interactive user experience with anything but the simplest graphics content.

So, how can this be implemented effectively in a Virtual Machine environment? To date, server-based GPU acceleration has come in two basic flavors: GPU Sharing and GPU Pass-through.

GPU Sharing: Scalability without the performance limitations

GPU Sharing relies on VDI software to provide a layer of abstraction that lets the client application behave as though it has its own physical, dedicated GPU, while the server's GPU (and driver) can think it's responding to one master host. The VDI hypervisor running on the server intercepts API calls and translates commands, drawing contexts, and process-specific address spaces, before passing along to the graphics driver.

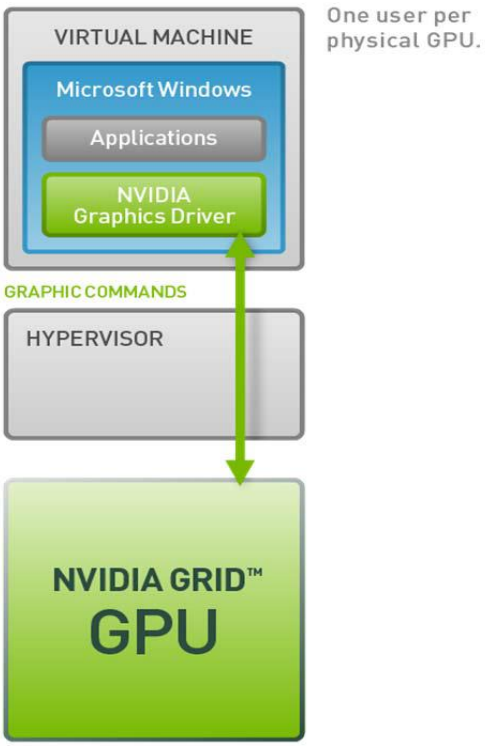


GPU Sharing is a reasonable solution for many, but not an ideal solution for all. It can perform effectively with simple applications and visuals and support concurrent users (CCUs), but the extensive compute cycles spent abstracting complex 3D rendering will add latency and reduce performance. Furthermore, the reliance on API translation means 100% application compatibility is impossible to guarantee. For example, applications which leverage features from the most recent OpenGL versions may not run as expected.

GPU Pass-through: Performance for Designers and Power Users

So if the software overhead of GPU Sharing is a problem, then why not go ahead and actually dedicate one physical GPU in the server to each hosted client? Well, that's precisely how systems are configured in servers implementing GPU Pass-through.

Unlike the rest of the physical system components, which are represented as multiple virtual instances to multiple clients by the hypervisor, the Pass-through GPU is not abstracted at all, but remains one physical device. Each hosted virtual machine gets its own dedicated GPU, eliminating the software abstraction and the performance penalty that goes with it. For example, a VDI server with 2 NVIDIA GRID K1 boards (4 GPUs per board) can support 8 simultaneous users.

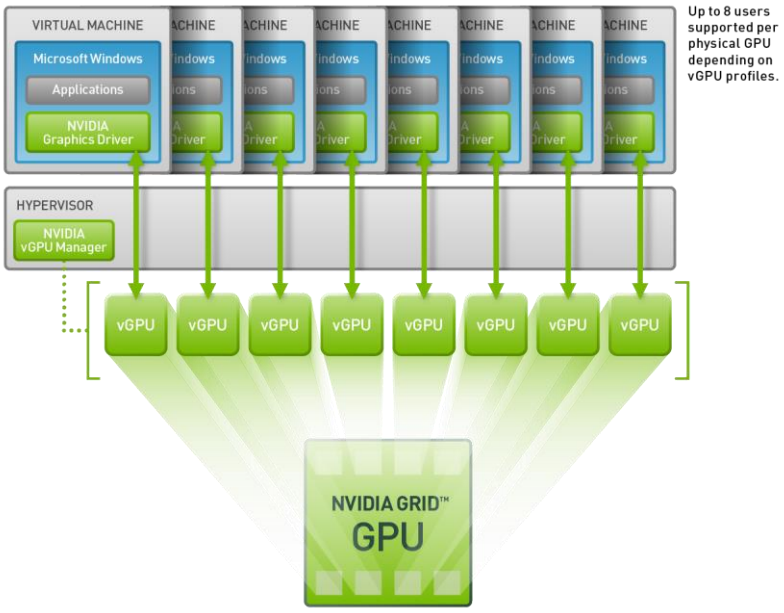


GPU Pass-through can make a lot of sense serving the power user, whose need for performance already demands a dedicated GPU. Consider the mechanical engineer who demands no-compromise visuals and for whom 100 GB file transfers are a daily occurrence. Instead of a Quadro board at his desk, he or she is tapping the power of a GRID GPU in the server, and reaping the benefits in data security, remote visualization, and support for BYOD hardware.

The vGPU Benefit

The inclusion of **vGPU™** support in Citrix 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.

In a Citrix environment, every VM communicates through XenDesktop to its own dedicated vGPU driver, for which one instance exists per VM. Each vGPU driver sends command and control to the one physical GPU, using its own dedicated input channel. As frames are rendered, the driver returns rendered frames back to the virtual desktop, which then streams it back to the remote host.

For VMWare users, GRID solutions are fully interoperable with from third-party vendors like Teradici, that capture rendered window images, encode them via PCoIP, and stream the window content to the client for subsequent decode.

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 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.

The rack server platform for the Dell Wyse Datacenter solution is the best-in-class Dell PowerEdge R720 (12G). This dual socket CPU platform runs the fastest Intel Xeon E5-2600 family of processors, can host up to 768GB RAM, and supports up to 16 2.5" SAS disks. The graphics-enabled Dell PowerEdge R720 offers uncompromising performance and scalability in a 2U form factor. For more information please visit: Link

A typical server configuration for a graphics enabled VDI solution is shown below:

Local Tier 1 Graphics Compute Host	Local Tier 1 Management Host
2 x Intel Xeon E5-2680v2 Processor (2.8Ghz)	2 x Intel Xeon E5-2670v2 Processor (2.5Ghz)
96GB Memory (6 x 16GB DIMMs @ 1600Mhz)	96GB Memory (6 x 16GB DIMMs @ 1600Mhz)
Citrix XenServer on 12 x 300GB 15K SAS disks	Citrix XenServer on 2 x 300GB 15K SAS disks
PERC H710 Integrated RAID Controller – RAID10	Broadcom 5720 1Gb QP NDC (LAN/iSCSI)
Broadcom 5720 1Gb QP NDC (LAN)	Broadcom 5719 1Gb QP NIC (LAN/iSCSI)
Broadcom 5720 1Gb DP NIC (LAN)	iDRAC7 Enterprise w/ vFlash, 8GB SD
iDRAC7 Enterprise w/ vFlash, 8GB SD	2 x 750W PSUs
2 x 1100W PSUs	

In the above configurations, the R720-based Dell Wyse Datacenter Solution can support the following user counts per server based on the configurations specified. vSGA and vDGA numbers are with VMware vSphere while vGPU numbers were obtained with Citrix XenServer as it is the only supported hypervisor currently. For more information on this graphics solution, please refer to the Dell Wyse Datacenter graphics reference architecture available here.

Local/ Shared Tier 1 Rack Densities		
Mode	**NVidia K1**	**NVidia K2**
vSGA (Shared)	18	20
vDGA (Pass-Through)	8	4
vGPU	32 (K140Q)	8 (K260Q)

Acknowledgements

Alex Herrera & the NVIDIA team for all their help and guidance with NVIDIA graphics cards.

Peter Fine & the Dell Wyse Solution Engineering team for their guidance with Dell hardware.

About the Author

Manish Chacko is a Sr. Technical Marketing Advisor for Citrix-based solutions at Dell. Before writing about technology, Manish has spent time designing, implementing and supporting technology- in IT, Systems Engineering & Network Performance/Monitoring. Manish has been a long-time Dell customer & Advocate before becoming a Dell employee.