WHITEPAPER | JUNE 2015

NVIDIA GRID vGPU: DELIVERING SCALABLE GRAPHICS-RICH VIRTUAL DESKTOPS

by Alex Herrera, Senior Analyst

NVIDIA

Table of Contents

EXECUTIVE SUMMARY

NVIDIA's GRID technology delivers a high-performance, interactive visual experience remotely, making complex 2D and 3D content accessible anywhere, any time, on any device. With GRID boards in a server, virtual desktop clients can for the first time access rich visual content with the largest datasets, in high resolution with interactive performance. They can, whether they're at the office, on the road, or off the clock checking in via a laptop, tablet or even phone.

Running on GRID-enabled servers supplied by leading OEMs, and supported by the leading virtualization solutions from VMware and Citrix, NVIDIA GRID vGPU (virtual GPU) technology delivers desktop-class performance that scales gracefully with a multitude of clients. Harnessing NVIDIA's GRID vGPU technology, professional-caliber visual computing is now ready for a shift to datacenters and clouds, to whatever degree meets the needs of the business.

Figure 1: GRID virtual GPU technology: server-side rendering of rich 3D content, delivered wherever, whenever.

THE NEW FRONTIER: VIRTUALIZED GPUS IN THE CLOUD

Today, server-centric computing is back capturing mindshare in a big way. It has a different shape than in it has in the past, as well as a variety of names and contexts—cloud computing, virtual desktop infrastructure (VDI), client consolidation infrastructure (CCI) and hosted virtual desktops (HVD). But in essence all imply the same basic idea: moving the heavy computation and data to a central resource, accessible by many rather than one.

Unfortunately, one important computing segment has largely been left behind in this transition from client desktops to virtualized clouds and datacenters. Historical constraints in both available technology and network infrastructure have conspired to preclude applications that demand anything but the most basic level of graphics capability. But that's finally changing, as infrastructure capabilities have matured, and now, the ability to remotely render and deliver high-performance graphics desktops has arrived.

When it comes to visualization, it's time for many to rethink the old paradigm of copying models and data from server to client to keep pixel bandwidth local, and consider a new one: leaving big data in a central computing resources, rendering visuals on a server with GRID hardware, distributing the graphics workload among GPUs with vGPU technology, and shipping only the resulting pixels to clients.

The reason is simple: a quality interactive visual experience always matters, and delivering on that experience always requires a GPU. That's a premise validated throughout the evolution of personal computing. And it's a premise that's just as valid for any enterprise's IT plans, whether those plans focus more on GPU-equipped clients like PCs and workstations, or whether it's a future that includes GPUs in the datacenter.

ACCESS ANYTIME, ANYWHERE, WITH CUSTOMIZED PERFORMANCE

In the era of global enterprises—operations, personnel, and supply chains—it's no longer viable to move the people with the necessary skills to a physical workspace, near the data. The reverse is now the goal, shaping computing environments and harnessing IT tools to instead move the workspace and data to the personnel that need it. A virtualized, centralized computing model is one such environment, and NVIDIA's GRID technology is one such tool.

With a GRID-enabled virtualized environment, team members don't need to share the same physical location. Consider the needs of a modern global architectural firm, where the architect, client, building contractor, and construction site might all reside in separate cities, countries, or even continents. With remote computing environments enabled by GRID technology, all can create, review and collaborate with rich, high-fidelity 3D visuals, wherever they happen to be. Keeping visual data in one place not only eliminates lengthy transfer delays, but makes for simpler and more robust development, without the added complication of multiple, extraneous copies of data scattered around the enterprise. Add vGPU technology and you get all this, but with greater economies, both in hardware utilization through GPU-sharing and in lower TCO through scaling.

Virtual desktop hosting not only makes access location and time independent, it allows administrators to scale the visual performance of that access, individually customizing the experience based on the needs of the user. For example, a designer or animator creating 3D content might be allocated all of a server's physical GPU resources for maximum productivity. By contrast, a power user—who may not be actively creating content, but needs to quickly, accurately view and mark up project material—might be allocated just a half or a quarter of those resources. Consider the daily needs of the product marketer, sales engineer, and support technician. Similarly, knowledge workers working primarily with 2D applications and GUIs, or viewing static 3D images might have their needs satisfied with only a small fraction of the host machine's overall physical resources.

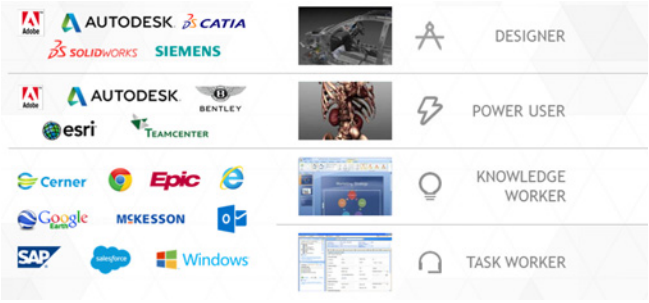

Figure 2: Project complexity and aggressive schedules demand timely accurate access to visual data by all.

EASING BIG DATA'S PRESSURE ON I.T.

Now, one could argue that anytime/anywhere computing can be accomplished with a conventional client-side computing model, without virtualized access to a central computing resource. To do so, an IT department and its users need to keep all the enterprise's clients—deskside and mobile, across the globe—up to date with the latest copies of project data. That's essentially what enterprises have had to do all along, and to some degree are still doing. However, even if IT personnel and infrastructure were willing to continue taking on that logistical challenge—tracking project development and ensuring each copy of client data remains coherent—the distributed, client-side model begins to fail in the age of exploding data.

Because when the data sets get big, copying them from where they're stored to where they're needed can waste minutes or hours, time businesses can ill afford in an ever more competitive climate. Consider that even a relatively small 100 MB CAD or AEC file can chew up several minutes of lag time, while transferring a short 4K video sequence can consume hours. For more and more applications, leaving the data where it is, and instead shipping the visual representation of that data—a pixel stream—can be far less taxing.

SECURITY

Given the magnitudes of both the investment and revenue at stake, protecting IP and digital content has justifiably emerged as a top IT priority across the business world. And for many, the single most compelling feature of a centralized computing topology is security. When content is both created and stored in the cloud, source data need never leave the prescribed, protected boundaries of the enterprise.

Even the pixels that do stream across the WAN are secure, as desktop delivery protocols from providers like Citrix and VMware provide a variety of customizable, government-level security protocols and cryptography. With a centralized, GRID-enabled computing environment, leaving a company laptop behind at the coffee shop becomes a minor inconvenience, not a monumental breach in security.

BYOD: THE ENCOURAGEMENT—OR ENCROACHMENT—OF HETEROGENEOUS CLIENTS IN THE ENTERPRISE

The incursion of smart digital devices into personal and working lives is undeniable. More and more, businesses need to deal with the increasing use of BYOD (Bring Your Own Device) in an effective and reasonable way. In the age of always-connected employees available 24/7, the line between personal and work devices has forever been blurred.

With the user's desktop OS, application and graphics running on the server, a client can take any number of forms, from thin-client, to conventional deskside or notebook Mac or PC, or even a smartphone or tablet. A virtualized GRID-enabled computing approach gives IT managers more options and more flexibility in supporting the unavoidable trend in BYOD, keeping users happy without throwing IT management practices into disarray.

Figure 3: Like it or not, all IT environments are going to have to deal with heterogeneous clients.

THE HISTORICAL STUMBLING BLOCKS TO HIGH-PERFORMANCE REMOTE GRAPHICS

With so many compelling advantages to anywhere/any-time/any-device remote visualization, graphics-rich VDI should be 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.

Historically, the simplest way to implement VDI has been through a fully-abstracted software implementation of a virtual machine, running on the server—call it the Soft PC. Without a GPU, the CPU has to process the entire graphics workload, a task for which CPUs are ill-equipped to handle, ultimately making the SoftPC model capable of little more than simple text-based applications with minimal graphics requirements. Rather, and for the same reason GPUs grew

to ubiquity on client platforms, a GPU becomes an absolute necessity to deliver visual performance in a server-hosted environment.

To date, server-side GPU acceleration has come in two basic flavors: Software GPU Sharing and GPU Pass-through. Software GPU sharing relies on host virtualization 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. Software running on the host server intercepts API calls and translates commands, drawing contexts, and process-specific address spaces, before passing along to the graphics driver.

Software GPU Sharing can accomplish what its name suggests—allowing multiple virtual machines to tap the rendering horsepower of a single, physical GPU—but it's burdened with two unfortunate tradeoffs: lower performance and limited application compatibility. It can perform effectively with simple applications and visuals, but the extensive compute cycles spent abstracting complex 3D rendering both adds latency and reduces performance. Furthermore, the reliance on API translation makes 100% application compatibility impossible to guarantee, as applications which leverage features from the most recent OpenGL versions, for example, may not run as expected.

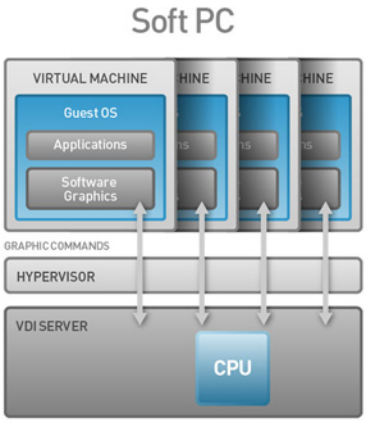

Figure 4: The Soft PC implementation for VDI: no GPU means no rich visual content.

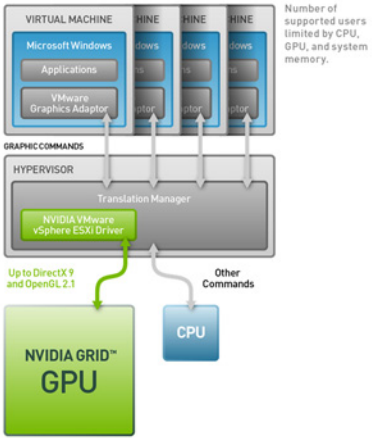

Figure 5: Software GPU sharing: supports multiple VMs, but compromises graphics performance

Given the performance penalties of sharing a GPU via software API intercept, virtualized deployments that needed higher and more dependable levels of performance have opted instead for an approach called GPU Pass-through (or Dedicated GPU). 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 allocated its own dedicated GPU, eliminating the software abstraction and the performance penalty that goes with it.

Ultimately, however, the primary advantage of GPU pass-through becomes its critical weakness. Tying a physical GPU to a single virtual machine limits the ability to leverage common server resources and goes against one of the main reasons to pursue a virtualized solution in the first place.

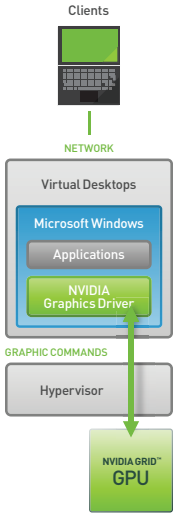

Figure 6: GPU Pass-Through: reasonable performance, but with no ability to share GPU among multiple clients

A NEW PARADIGM

GRID vGPU LETS REMOTE SERVERS DELIVER RICH, INTERACTIVE VISUAL CONTENT TO MULTIPLE CONCURRENT USERS

As a means to enable early adoption and stoke the fires for further development, GPU Sharing and GPU Pass-through provided workable means toward the ultimate end of delivering quality graphical experiences remotely. But neither approach is ideal—software sharing via API intercept constrains performance and limits compatibility, while dedicating a GPU via pass-through limits flexibility and optimal resource utilization.

Figure 7: All past server-side GPU approaches have their tradeoffs

What's ultimately needed is a single, unified approach that can deliver maximum throughput for a single user, backed by the guarantee of dedicated GPU resources. But it must be one that can also morph at the discretion of IT management—to seamlessly, gracefully scale to support a multitude of clients, without sacrificing performance or API compatibility due to software processing overhead. And ideally, it should be one that can be centrally managed, with one common, streamlined console to control it all. Enter NVIDIA's GRID vGPU technology.

PATENT-PENDING GRID vGPU TECHNOLOGY DELIVERS A LOW-LATENCY, HIGH-PERFORMANCE VDI EXPERIENCE

With GRID vGPU technology, NVIDIA engineers virtualized the GPU in hardware, allowing multiple virtual machines to share one physical GPU without the need for software hand-holding and API abstraction. GRID GPUs integrate patented functionality in silicon to provide virtualization in hardware: first and foremost a memory management unit (MMU) and dedicated per-VM input buffers. The GRID GPU's Memory Management Unit (MMU) allocates, maps and translates a VM's virtual address to the host server's physical address, allowing each VM to works in its own address space without interfering with the others.

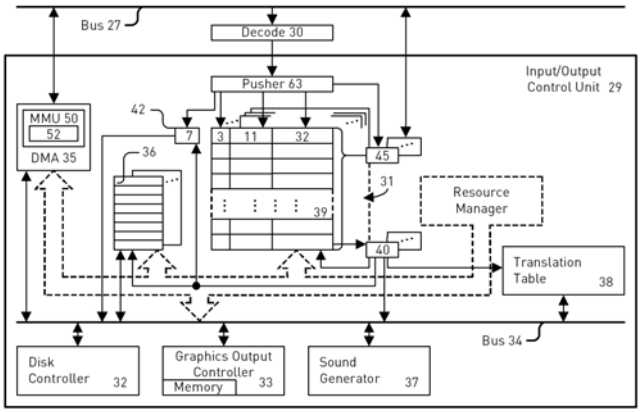

Figure 8: Beginning with the Kepler generation, a GRID GPU's MMU and multi-channel input buffers facilitate true hardware-virtualizable graphics

Working hand in hand with the MMU are hundreds of independent input buffers, each dedicated to a different VM, isolating its graphics stream within independent rendering contexts. The combination of the address-space unifying MMU and VM-specific input buffers forms the linchpin to deliver the industry's first truly hardware-virtualizable GPU.

FLEXIBLE, HIGH-PERFORMANCE GPU RESOURCE ALLOCATION

With virtualization support embedded directly in silicon, GRID vGPU alleviates the necessity of software abstraction to share GPU resources, eliminating performance-robbing CPU overhead, as well as concerns over application reliability and compatibility. Native GRID vGPU drivers leverage the same code base of NVIDIA Quadro GPUs, built on years of development and testing and running hundreds of professional applications in millions of PCs and ISV-certified workstations around the world.

GRID vGPU technology facilitates optimal GPU sharing, but it's another matter to ensure that each hosted virtual machine—and each user that machine represents—can get

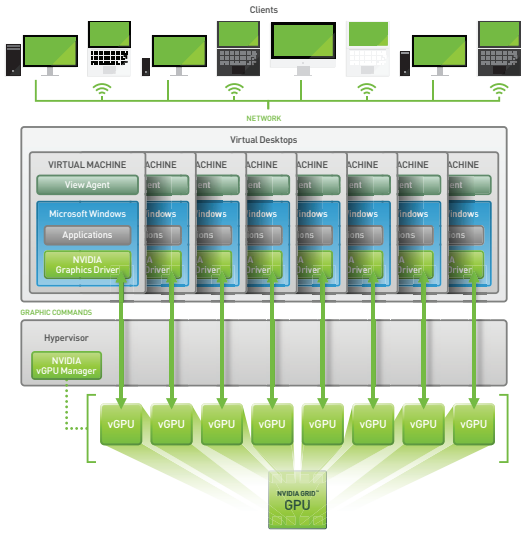

Figure 9: All virtual machines share one physical GPU with GRID vGPU's hardware-based virtualization

enough GPU resources to effectively process that machine's workloads timely and effectively. With GRID vGPU, IT administrators balance the graphics demand of users with the physical resources available via vGPU profiles. Each profile specifies how much memory and, on average, what fraction of a GRID GPU total available processing power that virtual machine can count on. IT administrators can ensure each user is adequately provisioned simply by selecting the appropriate profile.

A system manager might share one GRID GPU with four power users, while allocating another to eight knowledge workers with lower visual processing demands. And what about the hardest-core designer or engineer with the highest graphic demands, demands that would have

previously been served by the dedicated, GPU Pass-through model? With GRID vGPU technology, provisioning an entire physical GPU to one client requires no special-casing, but instead simply means choosing the proper GRID profile— for example, selecting the GRID K280Q profile dedicates all of a GRID K2 GPU's resources to a single VM.

Use Case	Virtual GPU			Per-VM Resources		
	vGPU Profile	# per GPU*	# per Board**	CUDA Cores (time-slice shared)	Frame Buffer (Dedicated)	Max Displays (2560x1600)
Designer	GRID K280Q	1	2	1536	4GB	4
Designer	GRID K260Q	2	4	1536	2GB	4
Designer/Power User	GRID K240Q	4	8	1536	1GB	2
Designer/Power User	GRID K220Q	8	16	1536	512MB	2
Entry Designer	GRID K180Q	1	4	192	4GB	4
Power User	GRID K160Q	2	8	192	2GB	4
Power User	GRID K140Q	4	16	192	1GB	2
Power User	GRID K120Q	8	32	192	512MB	2

Table 1: Today's snapshot of GRID's evolving set of vGPU profiles.

GRID vGPU SIMPLIFIES IT MANAGEMENT

With its seamless, profile-driven approach, GRID vGPU technology makes it easy to manage resources and serve modern IT demands—demands that are constantly in flux, thanks to a work force that's increasingly scattered and transient, and a work flow that's more dynamic than ever. Need to quickly scale resources as staff transitions in and out over the life of the project? An administrator can quickly re-provision GRID hardware with different profiles, and what had been a dedicated graphics resource for a single user quickly and cleanly becomes a resource shared among several—or vice versa.

And what if new hires or consultants aren't located on campus or even on the same continent? With a traditional, client-side workstation strategy, it might be time to burn extra time and money getting an IT professional on a plane over to the remote site to manage the installation. But with high-performance graphical VDI, enabled by GRID vGPU technology, provisioning new virtual desktops happens in the datacenter—controlled via management console—and the distant user does little more than plug in a thin-client and connect a monitor.

IT staff reap the benefits of a centralized, datacenter-driven computing approach not just for installs, but for managing computing resources over the life of the project. With GRID, an administrator—in one place leveraging one, unified management console—can handle most patches, updates, and other maintenance, regardless of where the end-users and their clients reside.

THE GRID ECOSYSTEM

The complete GRID solution consists of three primary components: GRID GPUs (and software) provided by NVIDIA, GRID-certified servers, and GRID-compatible virtualization software. Supporting server OEMs include Dell, HP, Lenovo, Cisco, and Supermicro. And GRID vGPU is now supported by the industry's leading virtualization platforms, including VMware Horizon and vSphere, and Citrix XenApp, XenDesktop, and XenServer solutions.

For everything you need to know about GRID and its supporting partners, visit www.nvidia.com/vdi. And for a complete, up-to-date listing of GRID-certified applications, check out www.nvidia.com/gridcertifications.

Figure 10: GRID certified platforms from industry partners

GRID AND vGPU: ALL THE TECHNOLOGY NEEDED TO VIRTUALIZE ENTERPRISE COMPUTING

Today, GPUs are found in virtually every computing device we interact with at work, play, or leisure. Rendering graphics with a desktop PC or workstation GPU is a sensible paradigm that has withstood the test of time and that continues to serve a vital role in the enterprise. But it's no longer the only paradigm. Because with GRID, NVIDIA is making the GPU location-agnostic, promising interactive, high-performance 3D graphics delivered from clouds and datacenters—for anyone, at any time, on any device.

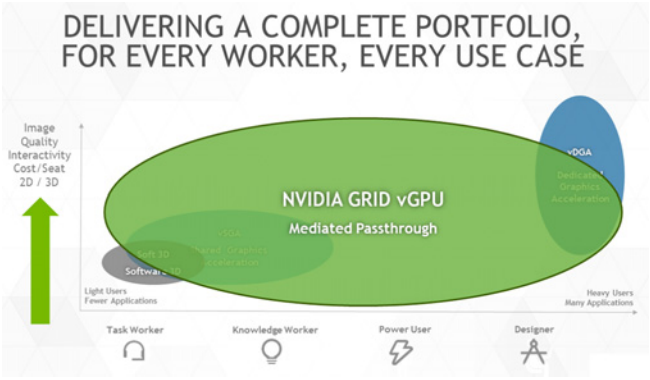

Figure 11: GRID vGPU offers the best of both worlds: optimal graphics performance and compatibility that scales by client

With vGPU technology, GRID achieves the same ends as existing remote graphics delivery models—software-based GPU sharing and GPU pass-through—but improves significantly upon their means. Employing patent-pending hardware in silicon, NVIDIA GRID vGPU shares graphics resources elegantly and effectively, with minimal processing overhead and maximum application compatibility. And with GRID vGPU, there's no more special-case choices of sharing or dedicating GPU resources. There's one mode, one management console, and one seamless way to allocate and manage graphics resources for one client or many clients.

Addressing many of the thorniest IT issues facing businesses today, a centralized, virtualized computing environment offers compelling benefits in security, IT management, managing big data, and addressing the onslaught of BYOD in the enterprise. With GRID and vGPU technology, that environment and those benefits are now available to enterprise users and applications demanding interactive and high-performance visual computing.

To learn more about NVIDIA GRID visit
www.nvidia.com/vdi