NVIDIA VGX:
Enabling the Enterprise VDI and Cloud Gaming Market

Executive Summary

NVIDIA today launched one of the most significant initiatives in the history of the company that, if delivered as promised, could propel them into a position as a top enterprise technology player. Called “NVIDIA VGX”, NVIDIA strives to remove many of the remaining barriers that have stood in the way of full cloud VDI implementations, and doing this for some of the most challenging and monetizable cloud workloads. NVIDIA and their partners are attacking the scalability, latency, and performance per watt challenges that had previously hindered large scale rollouts of remote desktop installations for designers and power users. They are doing this with a first of its kind and highly patented hardware GPU virtualization, lower latency remote display architecture, and high performance per watt Kepler GPUs. NVIDIA has launched with the support from all the key server OEMs, hypervisor providers, and cloud gaming middleware providers. If NVIDIA and its partners can successfully execute, they will have enabled the first true delivery of anywhere, anytime computing, for any application, a goal of the industry for the past 20 years.

Anywhere, Anytime Computing Vision

Over 20 years ago in early 1990’s, companies were declaring the need to enable anywhere, anytime computing and communications. This was an early take on "continuous computing", where users would interact with their content anywhere on a myriad of different devices from tablets, to phones, to PCs to cars to kitchen countertops. The industry has been grappling with many challenges to deliver this, including security, cost, content protection, bandwidth and overall user experience. The cost of displays up until now had made an acceptable, portable experience very expensive and out of reach of many. Until capacitive touch, users required a suboptimal keyboard and trackpad or a mouse to get a decent tablet or phone experience. Even inside the home, until Wi-Fi N, a wireless experience was unacceptable for video and moving files around between multiple devices. One of the last remaining issues to address the vision is centered on the holistic user experience.

Today's User Challenges to Cloud Experiences

All successful experiences with computing devices have one thing in common: responsiveness. Humans have different expectations depending on what they are doing. Just look at the difference in our responsiveness expectations between two things many of us do every day; checking Facebook and clicking an icon on a PC desktop. We are fine waiting one or two seconds to see Facebook status but would be very unhappy if it took two seconds for something to happen after we clicked an icon. Even the hour-glass, while annoying, at least assures us that something is happening.

Today's most successful cloud applications deliver experiences that aren't real-time. Lagginess is accepted. Facebook doesn't require an instant update. Buying something on Amazon doesn't require an instantaneous reaction as users are OK with seconds in lag time. Users have almost instantaneous expectations when it comes to working with their own PC desktop and even while playing games. Imagine if after one second of swiping your finger across an item in Fruit Ninja, you actually saw and heard the cut. Unacceptable experiences are the primary reason these usage models haven't successfully made it to the cloud. When it comes to virtual desktops or cloud gaming, research from NVIDIA suggests that no more than a 200ms lag is acceptable to the end user. Before NVIDIA VGX, the technology infrastructure hadn't existed to deliver an acceptable performance.

Layman's Explanation of a Virtualized Cloud Experience

Virtualizing anything in the cloud is a very highly sophisticated endeavor. There have been billions invested in R&D, multiple patents and books written on the subject. That doesn't mean the layman cannot appreciate it, though.

Think of it like this: the computing is happening in the cloud and sending you a movie of what it looks like. It's like when you are watching a baseball game on TV; it isn't happening in your living room, it is taking place thousands of miles away but broadcast into your living room. When a Windows desktop is being virtualized in the cloud, the computing is taking place hundreds of miles away in a data center and the image are being broadcast onto your device. Cloud gaming is the same; the gaming hardware is thousands of miles away in a data center and the video of what is happening is sent to your device.

Today's Technical and Business Challenges to the Cloud

Windows desktops and cloud gaming both exist today, but it is very challenging technologically which leads to business challenges as well.

Scalability

Let's look at an extreme example first, cloud gaming. In cloud gaming, only *one user can be connected to one graphics card in the cloud at the same time.* You can imagine just how inefficient this is and how this impacts profitability. Your cloud gaming business scale is tied to the amount of graphics cards you can squeeze into your data center. Because the servers need be located within 100s of miles away, loads cannot be balanced across continents and time zones, ensuring idle servers.

Latency

Another issue is latency which is an issue in cloud gaming and VDI (Virtual Desktop Interface). Latency is the lag time between when the user does something and gets a response back. As mentioned above, 200ms and below is the goal here. This is a huge challenge, given so many variables are fixed, like the carriers, but most of the lag time can be explained in the architecture. Let's look at VDI as an example of how latency is injected into the system. Today's VDI actually emulates the GPU in software

and converts the graphical output to a video stream in software. Generally speaking, things done in software and emulated are much slower than operations done in hardware. Architecture is an issue which adds latency. Even converting the display to a video format for streaming over the network takes four laborious steps, which as you see later in NVIDIA's VGX architecture, is reduced significantly.

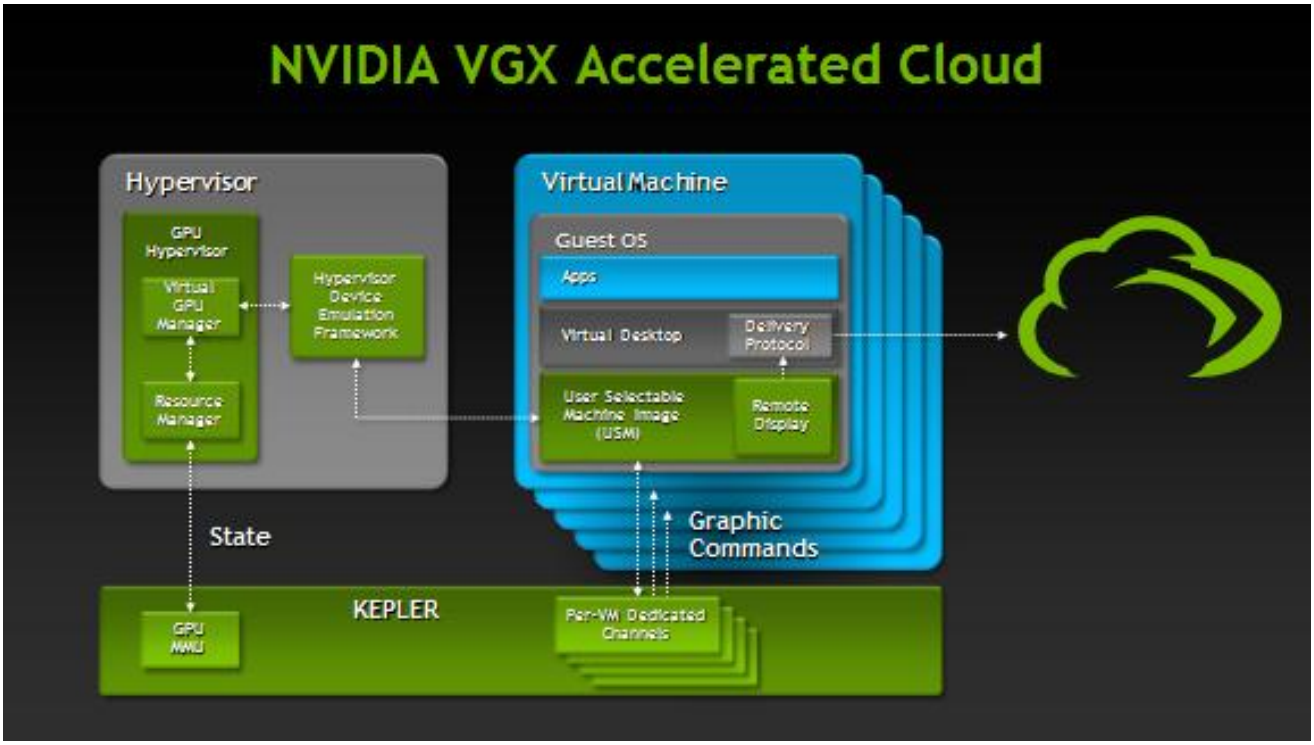
Latency in cloud gaming is addressed with some very elaborate data center solutions. Servers are placed geographically as close to users as possible to minimize latency. This is expensive, inefficient, and delivers an inconsistent quality of service. Servers are harder to leverage across large bases of users and cannot even assist in load balancing if users spike in a certain part of the country. The variability in the QOS (quality of service) also limits the potential target market that can be addressed. Entire continents today are ignored because of these challenges.

Performance per Watt

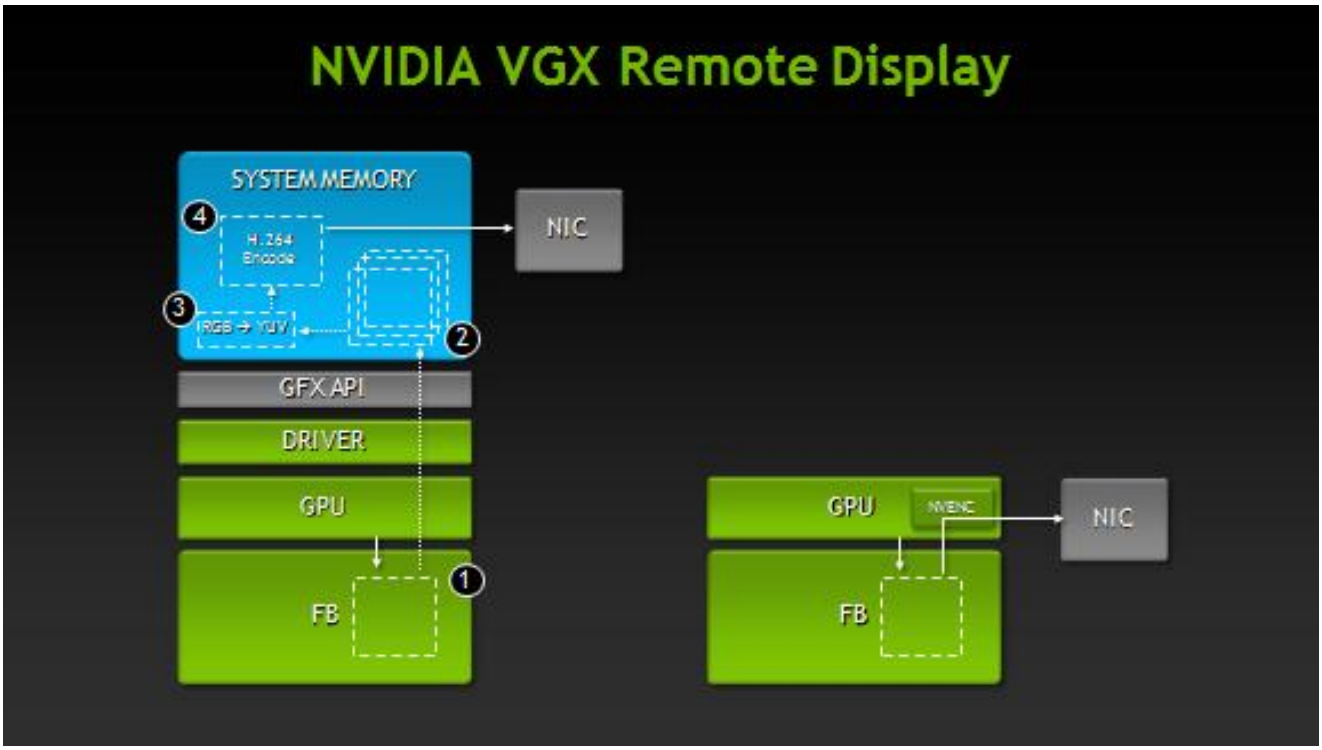
Today's enterprise VDI implementations like sharing a Windows desktop are done today in software. The challenge though is in the types of applications that can be done really well. Microsoft Word and PowerPoint is one thing, but designer applications like CATIA and Adobe CS6 require an altogether different level of performance per watt to virtualize remotely. Then there are classes of apps one notch below in terms of performance requirements for power users, who use the output of the designers. These are apps like PLM and medical imaging that require a substantial amount of computing and graphics performance. 3D games of course drive an immense amount of graphics and compute power that current VDI implementations simply cannot handle well.

How NVIDIA VGX Addresses the Challenge
NVIDIA is addressing all three of these barriers to cloud enablement of higher end applications for power users, designers, and all classes of cloud gamers.

NVIDIA VGX Virtualized GPU Solves Scalability Issues
Currently, GPUs cannot be shared in the cloud by different users. This has led to massive scalability issues for cloud gaming and virtualizing high end applications for designers and power users. NVIDIA's Kepler is the world's first GPU that can be virtualized in hardware, or shared, by many users in the cloud. Service providers can then install a few high-end NVIDIA Kepler-based VGX cards into servers and serve multiple users and application instances. VMware's Hypervisors and Citrix XenDesktop will both be supporting NVIDIA's VGX architecture.



NVIDIA VGX Remote Display Helps Solve Latency Issues
Currently, to remotely display graphics from the cloud to the end point is a very inefficient, multi-step and slow process. Data passes between the GPU, GPU RAM, through the graphics driver and graphics API into system memory and the CPU where it's converted into YUV then converted into an H.264 video stream and off to the end user. With NVIDIA VGX, the data passes between the GPU, GPU RAM and immediately converted into the H.264 video stream, *bypassing the GPU driver, graphics API, system memory, and the CPU*.



NVIDIA VGX Delivers Extensive Performance per Watt
Solving scalability and latency issues is a big part of the battle, but is incomplete without a significant bump in performance per watt. Kepler versus Fermi has doubled performance per watt and, all things equal, is now capable of servicing twice as many users, or servicing the same number of users with twice the complexity of a CAD design.

NVIDIA VGX Can Solve Real Business Problems Today

Great technology is one thing, but having that translate into real business value is sometimes different. Let me illustrate just a few use cases where NVIDIA VGX can help today.

Cloud Gaming Services

Game players can now immediately play any game available on virtually any piece of hardware they own, running any operating system and playing anywhere on the planet with decent connectivity. This translates into:

- Increased flexibility in data center location, saving money or increasing TAM (Total Available Market) and revenue
- Much more efficient player-GPU ratio, decreasing hardware cost, improving profits
- Improved game experience increasing penetration and repeat customers, increasing revenue
- Piracy-proofing titles for the publishers, increasing revenue.

NVIDIA GeForce Grid is the brand name that NVIDIA uses for cloud gaming since the GeForce gaming experience is extended to users via cloud computing, but it is based on the same GPU technology as NVIDIA VGX.

Enterprise Designer and Power User VDI

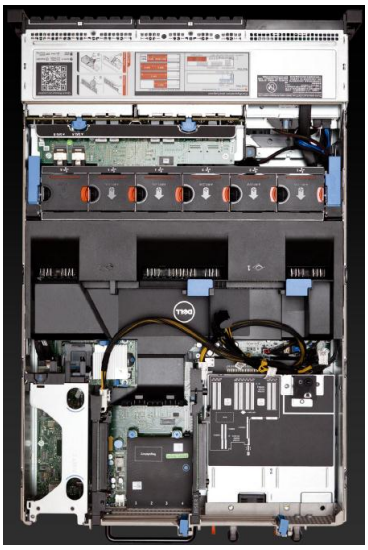
Enterprise designers and power users can now securely run the highest-end design applications on virtually any hardware they own running any operating system, anywhere on the planet with decent connectivity. This means:

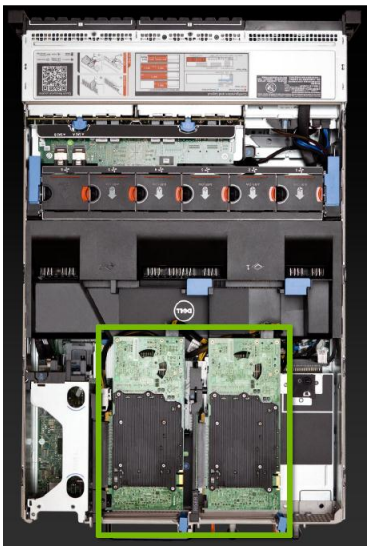
- Place-shifting and time-shifting on the highest performance apps, increasing productivity and employee satisfaction
- Real-time design and design review closer to the customer or decision maker, speeding time to market
- Much more secure environment as confidential data files never leave the premises

NVIDIA VGX Ecosystem Enablement

Virtualizing higher end applications in the enterprise and consumer space takes the support of many different types of players. NVIDIA VGX has a lot of support from the beginning, which is very unique for an initiative on this scale. This support must turn into action to be fully successful. The following ecosystem partners are supporting NVIDIA VGX at launch:

- *Server OEMS:* NVIDIA has the support of the largest server OEMs: HP, IBM, Dell, SuperMicro, Cisco, and Amazon. Any server that is certified to work with NVIDIA Tesla cards will physically work with the new NVIDIA VGX cards. Most rack mount servers have open PCI Express X16 slots and NVIDIA VGX cards will be placed in those open slots. *(See image below to see how NVIDIA VGX cards fit into server chassis.)*





- *Hypervisor:* NVIDIA has the support of Citrix, VMware, Microsoft and Xen hypervisors.
- *Cloud Gaming Middleware:* NVIDIA has the support from Gaikai, OTOY, G-cluster, Playcast, and Ubitus.

What is unique about this ecosystem is that they are all financially motivated to move ahead aggressively as it either means new revenues, increased profits, and/or lower costs.

Disruptive Nature of NVIDIA VGX and NVIDIA GeForce Grid

If it isn't apparent, NVIDIA VGX has the ability to disrupt many different markets as it rolls out into the future. I have listed just a few below:

- **Game consoles:** Every Smart TV, tablet and phone with decent video decoding and connectivity could become a game console and play every cloud enabled game. If users can do this, why would masses of gamers pay for a game console? Some will still choose consoles out of habit, lack of decent connectivity and the desire for the highest level of graphics, but many will not.
- **Physical retail games channel:** Cloud games virtually eliminate piracy and, if the game console market shrinks as outlined above, brick and mortar and e-tail outlets that once sold console games will reduce revenue significantly.
- **Business PC:** If users can access all my data and apps anywhere I am on any BYO (bring your own) device I already own, why does my enterprise need to buy and deploy more business PCs? Why not invest in an NVIDIA cloud and give employees their choice of devices, even one they bring from home, and allow them to securely connect and be productive? Many users and enterprises today will opt for local data access, but as mobile connectivity increases in reliability and speed and decreases in cost, more will opt for remote and mobile VDI solutions.

Conclusion

NVIDIA VGX has the ability to solve some of the last remaining issues that have kept the cloud from fulfilling its potential of truly an anytime, anywhere, any app experience. While applications like email, calendars and social media are very popular in the cloud, higher-order applications for designers, power users, and gamers are not. Current solutions do not provide sufficient scalability, low enough latency remote displays, or sufficient level of performance per watt to do any of these usage models well. NVIDIA VGX enables, for the very first time, hardware virtualized graphics, the lowest latency remote display, and high performance per watt capabilities to drive these usage models. VDI has been in somewhat of a holding pattern for years and NVIDIA VGX is just the disruptor that it needs to wake it from its slumber.

NVIDIA GeForce Grid is a disruptive technology that could change the face of the game console ecosystem and the business PC market as these devices become redundant in the long term view. Server OEMs, hypervisor vendors, and cloud gaming middleware vendors appear to be supportive of this initiative, well ahead of schedule compared to many other previous initiatives. If the NVIDIA VGX initiative can succeed, NVIDIA will have surely taken a coveted place at the technology enterprise table.

Author
Patrick Moorhead
President and Principal Analyst
Moor Insights & Strategy
www.MoorInsightsStrategy.com
Email: patrick@moorinsightsstrategy.com

Inquiries
Please contact us at the email address above if you would like to discuss this report and Moor Insights & Strategy will promptly respond.

Licensing

Disclosures
Moor Insights & Strategy has a consulting relationship with NVIDIA. No employees at the firm hold any equity positions with NVIDIA.