Building Great Virtual Desktops with NVIDIA Virtual GPU

How to prevent 15 common problems that can rob you of a great PC-like VDI experience

whitehat VIRTUAL TECHNOLOGIES

In 2006, Boeing discovered that the talent they needed to design and build their new 2.7 million-part Dreamliner 787 was not in the same city, the same country, or even all working for Boeing. Not only was this new airplane going to be pioneering, the way it would be built would be too.

Boeing partner, NVIDIA, would invent virtual GPU (vGPU) technology in a collaboration with Citrix and its own Citrix XenServer hypervisor, allowing hundreds of engineers from 45 different companies to work together, as one, in real time, on one of the most graphically demanding CAD/CAM/CAE applications in the world, Dassault Systèmes CATIA.

In March of 2013 this NVIDIA virtual GPU (vGPU) technology was released to the public, allowing virtual desktops to reach graphic-parity with physical PCs.

Seven years later, COVID-19 has forced the world to rethink how and where we work. NVIDIA vGPU adoption continues to increase, allowing today's web cams, conferencing tools, and training content delivery to remove any remaining geographic barriers, providing business with the capability to hire and position employees where they need to be to make the most difference for the business.

3

INCREASED GPU DEMAND

comparing Windows 7 and Windows 10

Google Chrome

↑ 36%

Firefox

↑ 59%

Microsoft Excel

↑ 53%

Microsoft Word

↑ 85%

Microsoft Powerpoint

↑ 64%

Skype

↑ 409%

Today every consumer computer that ships with an Intel Core processor includes an integrated GPU to process graphic content and application data, like web browsers, which all seek to take advantage of performance gains by leveraging the processor-native GPU capabilities. Intel's recently released Ice Lake (Gen 10) CPUs now come with up to 1.15 TFLOPS of compute performance and 64 GPU execution units allowing for monitor resolutions up to 8K. We can take from this that GPUs are no longer optional for even the most basic of computers to deliver modern applications, they are a near-requirement.

Yet, two of the biggest myths in the VDI world of Citrix Virtual Apps and Desktops or VMware Horizon, is that GPU is only necessary for fringe use cases where end users are working with graphic-intense applications such as CAD, although, **the number of computer applications using graphics accelerators for higher digital performance has doubled since 2012.**[1]

Web browsers such as Chrome, Edge and Firefox are some of the most resource-intensive applications in any environment and all of them can access GPU to improve page rendering times. Streaming video from YouTube or Vimeo, for example, must be compressed and decompressed before being delivered to the monitor for consumption successfully. GPU technology is responsible for smooth content delivery of streaming media.

With GPU, the Windows OS and graphics accelerated applications run faster and overall performance increases due to work being offloaded from the traditional processing engine in any computer, the CPU. The truth is, no matter how much may be claimed otherwise, GPU makes a significant difference in VDI environments **when NVIDIA virtual GPU and other technologies like it are setup properly.**

Application performance improves 36% with NVIDIA GPU

User productivity improves 13% with NVIDIA vGPU, or the equivalent of 260 additional productive hours per employee, per year

Help desks are 51% more efficient with GPU because end users are happier and not reporting issues as often

In a 2019 whitepaper, IDC research concluded that application performance improves 36% in virtual environments delivered with NVIDIA virtual GPU technology because applications were able to access increased processing capacity of the CPU + GPUs.[2]

Virtual desktop (VDI) user productivity improved 13% as applications improved in responsiveness, becoming more consistent and PC-like, meaning end users created fewer support tickets for problems they were experiencing. In fact, IDC found help desk operations were 51% more efficient in part because of the improved end user experience. The advantage of centralized management and the ability to include more graphic-intense use cases in the VDI environment reduced cost to support the endpoints by 36%, leading to an overall 49% reduction in TCO.[2]

Read that again, adding NVIDIA vGPU to VDI environments improved the end user experience, improved Help Desk efficiency by 51%, and cut the time it took to get the initial investment back, the ROI, by half. IDC's analysis revealed end users gained 36 productive hours per user, per year with NVIDIA vGPU. [2,3]

"...Participants (in the study) stressed how GRID vPC enabled better user experience for everyday applications such as Microsoft Office, Google Chrome, and PDF viewers. Interviewees also cited good support for the use of corporate-level video applications for training and other purposes".
-IDC[2]

A second pervasive myth is that implementing NVIDIA vGPU into a Citrix or VMware environment is as simple as installing some graphics cards and a bit of software. This view oversimplifies the effort involved significantly. Done correctly, incorporating NVIDIA vGPU into a Citrix Virtual Apps and Desktops or VMware Horizon environment will deliver a trifecta of benefits.

Cost reduction to support endpoints in the VDI environment

1. A radically improved end user experience.
2. Support costs per user should decrease, up to 50%. This number is backed up by IDC research.[2]
3. Achieving maximum user density for **knowledge workers** should yield an average cost of $4/user/month. This assumes a 1 GB vGPU profile over 3 years, cards + software.

An overall reduction in TCO

To achieve the trifecta you must: choose the right host and optimize it for NVIDIA vGPU, choose the right NVIDIA GPUs, optimize VM sizing, optimize for user density, be mindful of which endpoint OS and build you select, select the appropriate protocols for delivery of the desired content, and taking into account the capabilities of your desired endpoint(s), host servers, GPUs, hypervisor, virtualization platform and available bandwidth.

Improved Help Desk efficiency

Whitehat has been engaged 12+ times in 2020 so far specifically to help large organizations sort out implementation challenges or rescue struggling VDI projects entirely largely due to missing one or more of these decision points.

In an effort to assist the VMware Horizon and Citrix Virtual Apps and Desktop communities achieve a higher degree of success, Whitehat Virtual Technologies' engineers have pulled together a list of the most common challenges we see companies struggle with in their effort to deliver VDI projects on their own. What follows is a list of the most common challenges we see, where things can go sideways and what steps can be taken to avoid them on your road to delivering vGPU-enabled applications and desktops that deliver the trifecta of benefits, not the least of which is a true PC-like experience for end users.

#1

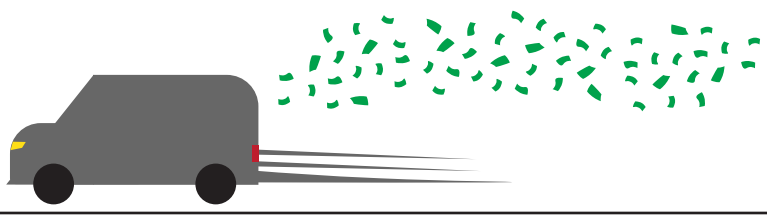
VDI Challenge #1:

The vGPU-enabled VMware or Citrix environment does not deliver the desired quality or end user experience when put in production, at scale.

The budget is spent, the experience is not a good one or does not scale well. X vendor said this would work, reality is saying something different. Maybe the system just needs to be tweaked? Right?

Root cause: Companies do not know there is a methodology for collecting accurate sizing data from their production environments for vGPU or have not developed one of their own. As a result, limited data, or no data and some guesstimates with results from a vendor calculator or two were used to determine the initial environment sizing that drove the purchasing decision.

Original Equipment Manufacturer (OEM) representatives and their resellers can only make sizing decisions based on the data they are given. Without proper data to drive environment sizing and purchase, price alone can become the most important factor in the decision. This can lead to making price driven design concessions without having the necessary data needed to understand the impact of those decisions. At the end of the day their objective is help you buy what you need, but their sizing responsibilities end with the purchase while the buyers responsibilities do not end until everything is successfully in production.



NVIDIA vGPU-enabled VMware Horizon or Citrix Virtual Apps and Desktops requires a bit of specialty knowledge. Vendors may not always be equipped with the knowledge or tools to master a VDI sizing exercise.

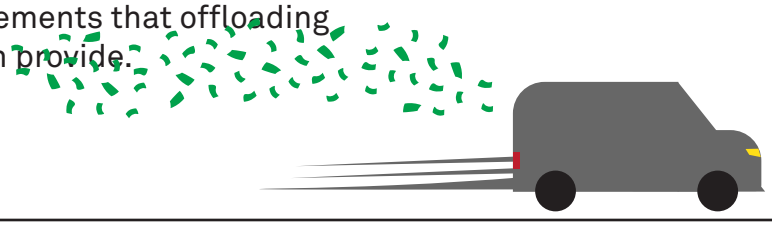
NVIDIA vGPU Best Practice: Identify each group of users with their unique application sets or work with a partner to complete a Desktop Transformation Assessment. Complete a proper Proof of Concept with all applications and a representative group of testers to establish how much CPU (GHz and cores), RAM and GPU are going to be needed to deliver an optimal end user experience. The end result should deliver specific sizing recommendations to give to your hardware vendors the supporting data to validate the claims

Mistake to Avoid: Making purchase decisions solely off a sizing calculator or using load estimation simulators with default out-of-the-box configurations instead of tuning the simulation to model your real environment-specific VDI behavior. A consultant may be required for this exercise, but done well, this effort will save tremendous pain downstream and a fair amount of budget.

Note: There are elements that simulators can't properly model. Login times are one example because they can't take pre-launches into account. Where possible test and size with real people and real data.

Speaking of load simulators, GPU-enabled servers typically achieve 16% up to 60%[3,4] additional user density per host achieved by offloading work from the CPUs and enabling them to handle more of the tasks they are better suited for. Without exception, we have never seen the addition of vGPU make server density worse. However, one very common simulator, when used out of the box with its default VDI simulation delivers results showing much lower density with vGPU than without it, artificially inflating cost per user numbers.

We can't emphasize this enough, optimizing user density must be part of the initial design when selecting CPU, RAM, the number of NVIDIA GPUs the host is able to physically support, and the desired end user vGPU profiles. Mistakes made in the design phase can easily destroy any user density improvements that offloading workloads from the CPU(s) to vGPU can provide.

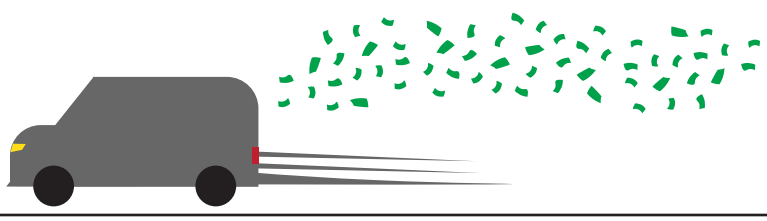


#2

VDI Challenge #2:

Beginning an NVIDIA vGPU project without having a solid understanding or any practical experience in how to build and manage an NVIDIA vGPU environment.

The design considerations for adding vGPU to Citrix Virtual Apps & Desktops or VMware Horizon requires very specific knowledge, ideally tempered with practical experience.



#3

VDI Challenge #3:

Forgetting to properly enable the NVIDIA GPUs.

When NVIDIA GPUs are not properly enabled there can be systemic issues that disrupt the overall user experience. Some common oversights include:

- Mismatching the driver versions of host and Operating System.
- Failing to properly configure profiles in terms of how they are carved out.
- Not properly pairing the NVIDIA GPUs with the proper AOS version in Nutanix HCI infrastructures and making sure that the CVMs are also enabled for NVIDIA vGPU use.
- GPO registry and optimization policies have inadvertently disabled hardware acceleration.
- Forgetting to set hardware PCIE buffers to align with the power use requirements of the NVIDIA GPUs.
- With VMware Horizon environments specifically, not making sure the template is configured to use the hardware, selecting the right mix of NVIDIA drivers and VMware ESX hypervisor versions.
- Not configuring the NVIDIA GPUs for NVIDIA Quadro Virtual Data Center Workstation (Quadro vDWS) or NVIDIA GRID vPC licenses, depending on the objective.
- With XenServer environments, not matching the right versions of XenServer hosts with the proper NVIDIA vGPU driver versions and enabling the host for GPU utilization.
- With Hyper-V environments, not realizing that NVIDIA GPUs can only be used in pass-through mode.
- Failing to properly configure the drivers and BIOS properly to enable the NVIDIA GPUs to function properly.
- Having an incompatible version of XenApp, XenDesktop, Citrix Virtual Apps and Desktops or VMware Horizon preventing the environment from working.
- Having policies set with the wrong virtual display adaptors to use NVIDIA vGPU as intended.

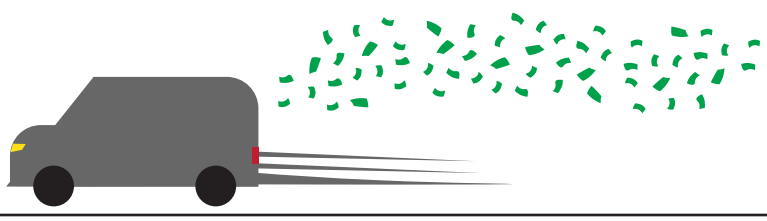
#4

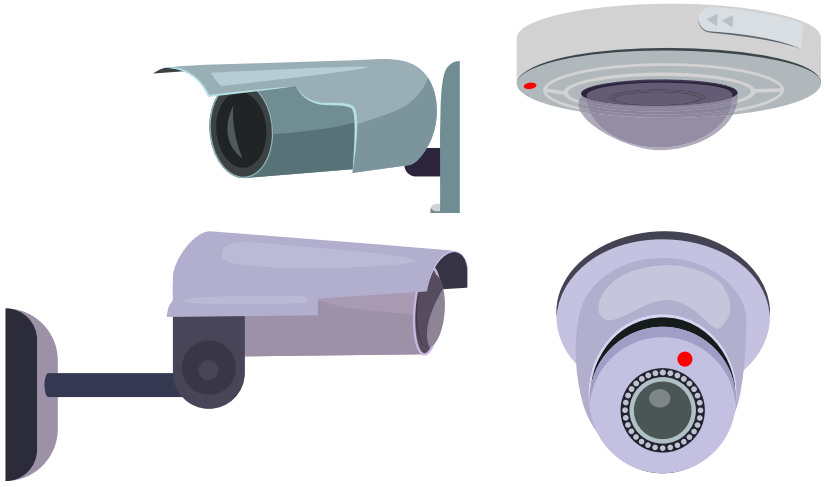
VDI Challenge #4:

Having PCoIP, Blast Extreme and Citrix HDX protocols misconfigured for use with NVIDIA vGPU.

PCoIP and Blast Extreme both have a strong use cases in a VMware Horizon environment. However, Blast Extreme provides benefits with NVIDIA vGPU in a VMware Horizon environment as it is the only one of the two protocols that can be decoded specifically by the GPU cards and handle traffic utilizing H.264 encoding.

Citrix provides multiple protocols to suit a variety of use cases including 3D Pro and Thinwire, Thinwire+ and DWM which leverage GPU capabilities in different ways.





#5

VDI Challenge #5:

Forgetting the importance of the host power settings.

When the host's power settings are not configured to properly leverage power and CPU frequencies at the BIOS level, performance and user density will be reduced resulting in an overall degraded experience. Power settings are not optimized for VDI by default from the manufacturer. Left unoptimized, host capacity will be wasted, and cost-per-user will creep higher unnecessarily.

#6

VDI Challenge #6:

Video playback failing to leverage UDP, negatively impacting video performance.

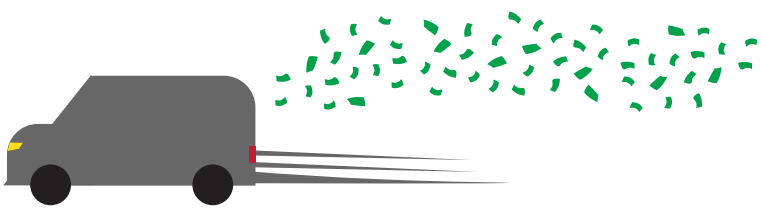
Protocols are worthy of their own topic entirely, however the most common challenge centers around picking the right protocol for video delivery.

TCP (Transmission Control Protocol) incorporates a data integrity checking and control element to confirm that packets arrive at their destination without errors. While great for reliability, the process adds overhead that can impact the quality of the content delivered. When looking at text or a picture, you want all of the data there. When watching a video, it is typically less important in most instances that every packet arrive as you are getting a new picture, or frame, several times a second.

UDP (User Datagram Protocol) does not have a built in quality control component or the overhead that goes along with it so it tends to be the protocol of choice for streaming video content where we are more concerned with moving large amounts of data to the endpoint than we are with an occasional error that might cause the screen to freeze or pixelate momentarily.

There are instances where TCP is preferred, but these are rare exceptions.

UDP is key for high quality media playback in all but the most bandwidth rich environments or in some extreme latency use cases. With substantial Internet bandwidth, TCP can deliver an acceptable experience, but it is not the protocol of choice in the normal bandwidth constrained environments we experience in the real world.

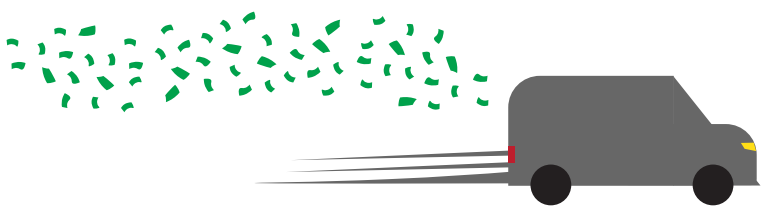


VDI Challenge #7:

The base Windows 10 image is not of sufficient quality to build an NVIDIA vGPU environment.

Anything less than a clean, optimized base Windows 10 image can negatively impact the production image and NVIDIA vGPU performance. Images imported from physical machines are often a problem as they carry drivers which are not suitable for virtualized environment. Additionally, Group policies (GPOs) can be at the root of the problem, being too numerous or in conflict with Citrix, VMware or NVIDIA vGPU configurations, making the overall experience inconsistent. A less than optimal base image will likely lead to a less than optimal build.

Note: VMware has created an incredibly powerful tool to help with OS optimization, appropriately named VMware OS Optimization Tool. While incredibly powerful, using it without a full understanding of its capabilities can have catastrophic effects. As an example, the most recent support incident we were engaged in involving a misconfigured VMware OS Optimization Tool fully disabled the organizations NVIDIA vGPU capabilities and made their gold image unusable. The tool is great, we just recommend developing custom templates for it, so you get all of the benefits without experiencing the potential downside.

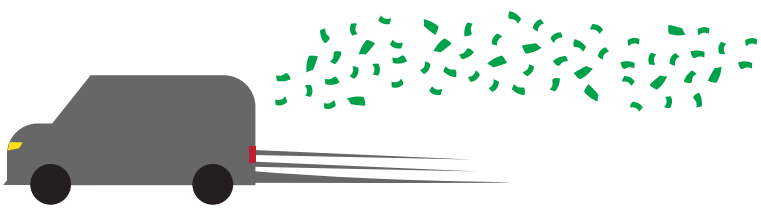


#8

VDI Challenge #8:

NVIDIA GPUs require NVIDIA virtual GPU software licensing, and it is common for this detail to be missed in the purchasing process.

For NVIDIA vGPU to work NVIDIA vGPU software licensing is a requirement. Coming up short, or not having any NVIDIA vGPU software licenses at all when ready to deploy, wastes project time and has sent many customers back to the well to request more budget to correct this oversight. It is an easy problem to avoid but is one that is quite common. This problem becomes exponentially more problematic when Disaster Recovery (DR) environments are added to the mix.



#9

VDI Challenge #9:

Underlying network issues undermine the quality of the NVIDIA vGPU environment.

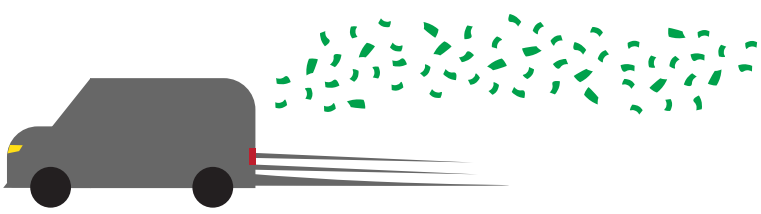
The greatest NVIDIA vGPU-accelerated desktops in the world are worthless if they can't get delivered across the network to the ultimate end user consistently. Focused efforts on getting NVIDIA vGPU projects off the ground can lead to missing the opportunity take the quality of the underlying network infrastructure into account, only to have these problems present themselves when the NVIDIA vGPU environment starts being taxed. It is important to evaluate the quality of the network before closing out the Proof of Concept phase. If the NVIDIA vGPU PoC is relatively small, network problems may be hidden by raw available capacity and masked in the PoC stage only to present at the worst possible time, under a production load most typically during a phased rollout.

#10

VDI Challenge #10:

Disagreements with the Security team and a lack of understanding on the unique characteristics around managing and mitigating risks associated with VDI environments frequently lead to delays or derail projects entirely.

There are subtleties in how NVIDIA vGPU environments work, where the risks are, and effective ways to mitigate them. Helping the security team understand the risks and how to mitigate these risks early can clear obstacles that can derail the NVIDIA vGPU project.

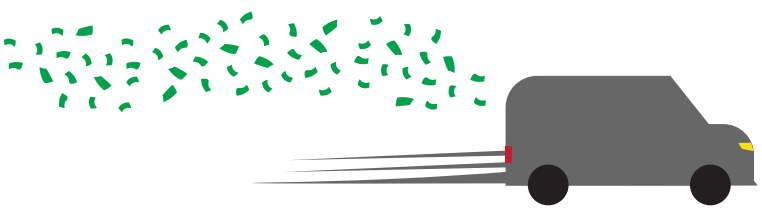


#11

VDI Challenge #11:

Lack of coordination across all teams required for success.

As pointed out in previous points, the success of any VDI environment can be impacted by seemingly unrelated IT systems and teams. In our experience, a full 60% of the VDI experience is dependent upon other IT systems working properly in concert to deliver an optimal user experience.



#12

VDI Challenge #12:

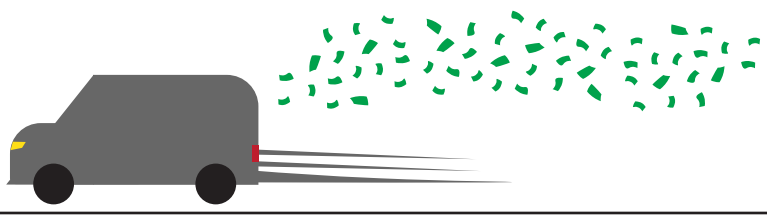
Not weighing the importance of server hardware form factor appropriately in the design phase.

Hardware brands can be a sort of religion unto themselves, but it would be a mistake to purchase any hardware without considering what impact the form factor/platform (HCI, Blade, traditional 3-tier-architecture, etc.) is going to have on the final design and the ultimate user density that the host can achieve. For a list of NVIDIA vGPU supported servers, see the hardware compatibility list at NVIDIA.com. The host server form factor/platform chosen can have a massive impact on maximizing user density and achieving lowest cost per user.

In full disclosure, Whitehat's own fully managed Titanium HCI VDI-in-a-Box appliance solution for both VMware and Citrix is built on a hyperconverged platform, but that does not necessarily make it the best choice in every application.

As an example, it is not uncommon to be able to load a server host with 70+/- VDI desktops sized to create a great end user experience, delivering the right balance of CPU and RAM to each user. However, when NVIDIA vGPU comes into the mix, the host may only be able to take one NVIDIA GPU supporting 16-32 users depending on the individual card selected. Maximum user density is no longer 70 VDI desktops per host, but 16 or possibly 32 VDI desktops, effectively "wasting" 50% of the hardware.

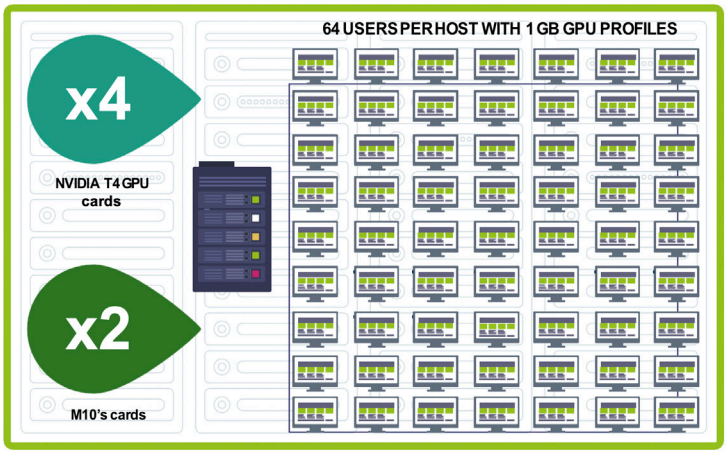
Note: One significant pain point centers around end user density for vGPU workloads. The number and type of NVIDIA GPUs that can be installed in host servers are often not seriously considered as part of the sizing effort. Failure here can render a portion of individual host server capacity wasted.



In practical terms you would likely fill this available capacity with non-GPU users which means another profile and added management complexity potentially cutting into the promised ROI numbers.

For maximum user density and support savings, the objective would be to align CPU, RAM and GPU to maximize the capacity of the host hardware and lower the cost per user.

Nutanix's latest release, the NX-3155G-G7 can hold up to four NVIDIA T4 Tensor Core GPUs or two NVIDIA M10 GPUs, capping user density at 64 users per host with 1 GB GPU profiles. This being the case, this hard cap could lead to changes in CPU selection or the amount of RAM configured for the host to try and keep all three metrics in line with one another.



NX-3155G-G7

Maximizing user density per host without sacrificing end user experience drives the cost per user down and the likelihood that an NVIDIA vGPU project will get off the ground successfully.

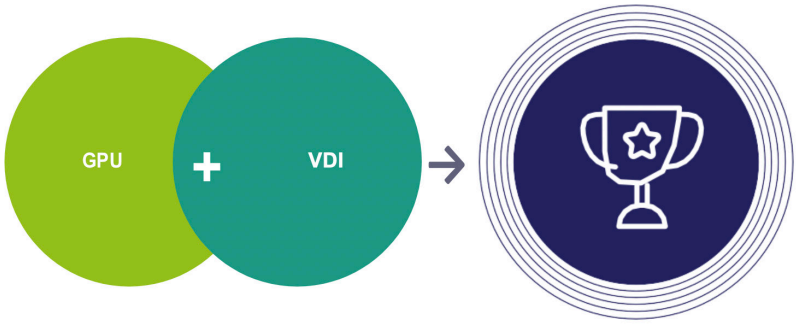
#13

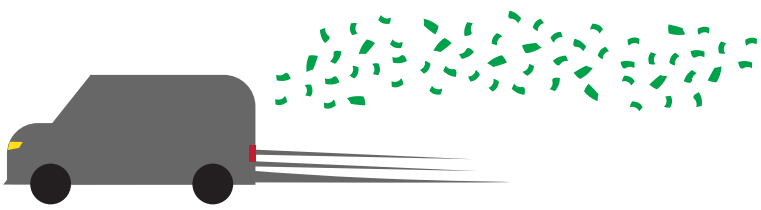
VDI Challenge #13:

Over-complicating the initial build.

Adding NVIDIA vGPUs to a VDI environment can have a dramatically positive effect on the end user and their willingness to embrace VDI technology. Mentioned indirectly in previous points, making sure Group Policies, the Windows 10 base image, and the network are all in good shape, number 13 here is this discussion at the macro level. It is easy to over complicate a build, the harder thing to do is to put forth a concentrated effort to keep the design and architecture of the NVIDIA vGPU environment as simple as possible.

Unnecessary added complexity undermines the potential for cost savings the infrastructure should deliver and often comes with a nice sidecar bolted to it of delivering a terrible end user experience.





#14

VDI Challenge #11:

Only realizing after the purchase that hardware for the NVIDIA vGPU project is undersized for the use case(s) after the budget dollars have been spent.

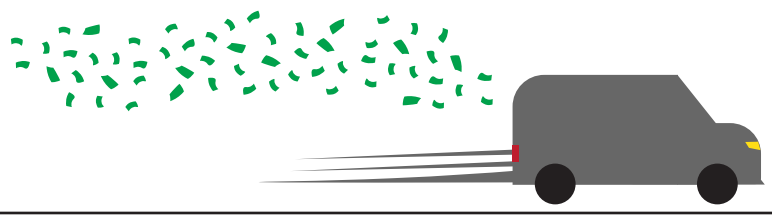
Our first rule of sizing hardware is to use real data from applications under load to size environments properly, do a Desktop Transformation Assessment or at least do a targeted assessment.

You don't buy Legos out of individual bins from a consultant and hope you bought all the right pieces for that Star Wars Imperial Star Destroyer you have been eyeing. You buy the kit knowing everything you need is in the box. The same can be said for NVIDIA vGPU environments (Star Wars Imperial Star Destroyer Kit not included).

Using guesstimates or simple calculators as the basis for purchasing host, SAN, and licensing before the real data is in, often does not end well. This problem is so big and pervasive that Whitehat developed a Citrix Apps & Desktops / Horizon View VDI-on-a-Box appliance just to help eliminate this problem.

NVIDIA vGPU implementations in this state can be a risk because of budgetary or political reasons that have nothing to do with the underlying technology which works for thousands of users around the world, every day.

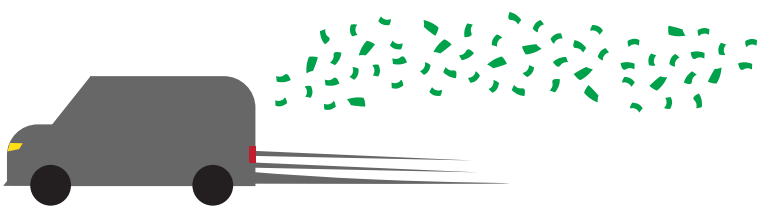
In the most common scenario, the budget has been spent and there is some pain around the fact that actual user density is not matching the vendor/partner projected user density.



Then the hard discussions start about reducing the scale of the project because the projected number of users cannot be accommodated or compromises in end user experience are evaluated to try and stretch this environment to cover the compute needs of a wider set of end users. Either the cost per user is going to be higher than expected or the end user experience is going to be less than expected. Neither is a good sign, but both can lead to the shelving of the project.

"We are now into the 4th generation of NVIDIA Virtual GPU technology. What worked in 2006, that the rest of the world got to experience in 2013, is even better in 2020. The question is not if it will work, the question is do you have the expertise to do the design and build, maximizing user experience and user density, while keeping cost-per-knowledge-worker or engineer in check."

Val King, CEO
Whitehat Virtual Technologies



#15

VDI Challenge #15:

Not recognizing the importance and impact protocols and codecs can have on end user experience and overall vGPU project cost and success.

How do we get content (Word documents, pictures, YouTube videos, webcam feeds, etc.) from the server to the person that wants to consume it? We use protocols and codecs.

Display protocols typically come in two parts. A server-side component and an endpoint-side component that lives on a laptop, thin client, etc. There is always at least one exception to the rule, and that is true here with HTML5 being a great example. HTML5 is a server-side only display protocol that leverages a web browser on the endpoint to complete its delivery.

Several elements can impact the choice of display protocols that need to be carefully considered in a given environment. Design considerations need to include the types of content being consumed, the protocols the endpoint can support, what the endpoints will be connected to, the protocols the virtualization platform can support (VMware Horizon or Citrix Virtual Apps and Desktops), required monitor resolution, internet bandwidth and latency, and LAN/WAN utilization.

If your experience with VDI or vGPU-enabled shared hosted desktops or VDI has not been a great one, it is likely due to one of the points above. Considering the COVID-19 world we live in now, Whitehat Virtual is offering free instances for a limited time to review vGPU deployments at a high level, correct what we can and point you in the right direction. This is something you should take advantage of if you are not having a great experience. The first step, as they say, is admitting you have a problem.

NVIDIA virtual GPU technology should improve the performance and end user experience of any Citrix Virtual Apps & Desktops or VMware Horizon environment. Whitehat can work with you to build an NVIDIA vGPU roadmap to provide guidance to take you from wherever you are to where you need to be or as tools to predict the end result of a VDI project before the budget is spent.

If you are ready to build your own NVIDIA vGPU roadmap for delivering any vGPU-enabled environment from heavy graphics-intense applications or simply trying to add Microsoft Teams, GoToMeeting or similar technologies and maybe some training video content on YouTube, **Whitehat can help.**

About Whitehat Virtual Technologies

Based in Austin, Texas with locations around the world, Whitehat Virtual Technologies was founded on the idea that work is something we do not necessarily a place we have to go. Employees, along with the businesses that invest in them, should expect to receive a great end user experience, enabling employees to be happier and more productive in what they do, while providing the business more IT flexibility, support savings, efficiency, and opportunities to scale and meet the needs of their customers.

Whitehat supports thousands of our customers' employees everyday by building, hosting, supporting, managing or co-managing their IT environments, frequently including Citrix, VMware and other cloud technologies.

Footnotes

[1] Data from Lakeside Software's SysTrack Community, 2017.

[2] IDC. (2019). NVIDIA Is Helping Organizations Provide Optimized Virtual Client Computing for Graphics and End-User Computing [White paper].

[3] NVIDIA. (2019) SEE THE DIFFERENCE FOR YOURSELF How to set up your own Windows 10 VDI test environment with NVIDIA Virtual GPU Solutions [White paper].

[4] Dell EMC. (2019) Quantifying the Impact of Virtual GPUs Benchmarking the User Experience in VMware Virtualized Environments with Dell EMC and NVIDIA nVector
August 2019 H17917

[5] Notebookcheck.net. (2020). Ice Lake: 10nm, fast GPU, and a new architecture. https://www.notebookcheck.net/Ice-Lake-Architecture-10-nm-Fast-GPU-and-Many-New-Features.427171.0.html

Business vector created by freepik and pch.vector - www.freepik.com

Disclaimer

This document is furnished "AS IS". Whitehat Virtual Technologies, LLC disclaims all warranties regarding the contents of this document, including, but not limited to, implied warranties of merchantability and fitness for any particular purpose. This document may contain technical or other inaccuracies or typographical errors.

This guide would not have been possible without:

Pablo Legorreta — Principal Enterprise Architect
Val King – CEO