TeamRGE
Remoting Graphics Experts

3D Graphics for Virtual Desktops Smackdown

Abstract

Understand 3D graphics for Virtual Desktops and Apps, receive unbiased and independent information from Dr. Bernhard Tritsch, Ruben Spruijt and Shawn Bass

THIS DOCUMENT IS PROVIDED "AS IS"
WITHOUT WARRANTY OF ANY KIND
FOR REFERENCE PURPOSES ONLY

3D Graphics for Virtual Desktops Smackdown

Version 2.12

Table of Contents

Introduction

*"Modern applications, window managers and browsers benefit from a GPU, which massively improves user experience. In the near future, **'GPU inside'** will be the standard for every Virtual Desktop solution."* **Dr. Benny Tritsch |** speaker, author, blogger

*"**GPUs are prevalent in practically every single endpoint device** that business consumers use on a daily basis. However, in virtual desktop environments the presence of a GPU is currently a very small percentage of desktops. Given the improvements that a GPU provides to a virtual desktop, it's not a question of whether or not you'll use GPUs in virtual desktops but rather when you'll use them."* **Shawn Bass** - EUC CTO of Desktop Technologies - VMware

*"**User Experience is king**. GPU capabilities and blazing fast solid state storage is the standard for a modern virtual workspace!"* **Ruben Spruijt**, CTO Atlantis Computing | MVP, CTP, vExpert

Are you looking for an independent overview of desktop virtualization solutions and are curious about different strategies? Are you interested in the use-cases and benefits of delivering 3D graphics desktops and applications remotely? Do you want detailed information about the features and functions each vendor is offering? Do you want to know how you can enable and leverage 3D graphics for virtual desktops? If so this is the whitepaper you definitely must read!

In the current market there is an increasing demand for unbiased information about hardware accelerated graphics for desktop virtualization solutions. This white paper is focused on solutions that are anticipated to have an important role in desktop virtualization and application remoting deployments. An overview of available features of the various solutions has been created to provide a better understanding of capabilities and to assist you in understanding important differences between these technologies.

Objectives

The goals of this whitepaper are to:

- Provide an application and desktop delivery solutions overview
- Explain different desktop virtualization concepts
- Explain the pros and cons of desktop virtualization
- Highlight the use-cases and benefits of 3D graphics for virtual desktops and applications
- Describe the vendors and their solutions to enable graphics within desktop virtualization
- Compare the features of the various 3D graphics for virtual desktops solutions

Intended Audience

This document is intended for IT Managers, architects, analysts, system administrators and IT Pros in general who are responsible for and/or interested in designing, implementing and maintaining hardware accelerated graphics desktop virtualization Infrastructures.

Vendor Involvement

All major vendors whose products are covered such as AMD, Citrix, HP, Intel, NVIDIA, Microsoft, Teradici and VMware have been approached in advance to create awareness of this whitepaper and discuss their solutions functionality and features.

Feedback

It is our intention to provide accurate, clear, complete and usable information. We appreciate your feedback. If you have any comments, corrections or suggestions for improvements to this document we want to hear from you! Please send an email to team@teamRGE.com include the product name, version number and the title of the document you're reading in your message. If you are searching for advice, design or implementation expertise? Let us know.

About

Acknowledgements

Bernhard Tritsch

Dr. Benny Tritsch is a business developer, principal consultant, market analyst, author, and all-around geek specializing in enterprise Windows remoting and virtualization solutions. He is Technical Director Central Europe at Lakeside Software and speaks around the world at several conferences each year, including Microsoft TechEd/Ignite, Citrix Synergy, VMware VMworld, BriForum and E2EVC. He has received the Microsoft Most Valuable Professional (MVP) award for RDS since 2004, the Citrix Technology Professional (CTP) since 2006 and the VMware vExpert in 2015. If you want to know more about Benny's activities in the virtualization community, check out his website at www.DrTritsch.com or follow @drtritsch on Twitter.

Shawn Bass

Shawn is a recognized expert in the End User Computing industry. With more than 20 years of consulting experience, he has worked with all facets of End User Computing (Terminal Services, Virtual Desktops, Application Virtualization and Enterprise Mobility Management) since the very beginning of these technologies. Shawn joined VMware in September 2014 to help lead and define a long-term strategy for VMware's End User Computing products and works closely with the product team in implementing the strategy in the years to come. He is currently serving as CTO of Desktop Technologies in the EUC business unit. Follow @shawnbass on Twitter or contact Shawn via email at shawn@shawnbass.com.

Ruben Spruijt

Ruben Spruijt is CTO at Atlantis Computing, responsible for driving vision, technology evangelism and thought leadership with Atlantis customers, partners and communities. Ruben is a well-regarded author, speaker, geek, market analyst, and all-around technologist. An established industry leader and luminary, he is one of only a few individuals in the world to hold three prestigious virtualization awards: Microsoft Most Valuable Professional (MVP), Citrix Technology Professional (CTP) and VMware vExpert. Ruben has presented more than 150 sessions at national and international events such as BriForum, Citrix iForum Japan, Citrix Synergy, Gartner Catalyst, Microsoft Ignite, Microsoft TechEd, NVIDIA GTC, and VMworld. Ruben co-founded several independent industry analysis bodies including ProjectVRC.team, Team Remote Graphics Experts (TeamRGE), AppVirtGURU, WhatMatrix. He has created and co-authored multiple disruptive 'Smackdown' research whitepapers. To contact Ruben directly send an email to ruben@rspruijt.com or follow Ruben on Twitter: @rspruijt.

Team Remoting Graphics Experts - TeamRGE

Shawn Bass, Bernhard Tritsch and Ruben Spruijt are the founders of TeamRGE. TeamRGE is a community group of experts with a focus on Remoting Graphics for Virtual Desktops and Applications. The goal of this group of thought leaders is to share unbiased and independent knowledge via blog posts, white papers, videos and presentations at local and international events.

Friends of TeamRGE

TeamWork - it's only through the effort and persistence of the 'Smackdown 'community team that we achieved the goals, a big thanks to these 'Friends of TeamRGE.'

Jits Langedijk

Jits Langedijk is Solution Architect at PQR. With over 12 years of experience in the End-User Computing space his main focus is on securely delivering remote applications, and desktops as well as access to data in the most flexible manner, any device and any location. Within the space of End-User Computing his primary focus is on solutions from Microsoft, Citrix and VMware. As a Solution Architect he helps organizations with their IT strategy, designs large and complex infrastructures and oversees the implementation as the technical project leader. Jits has presented at various (inter)national events about 3D Graphics (NVIDIA GTC 2015), Virtual Desktops Infrastructure (VDI/SBC) and Enterprise Mobility Management. Follow Jits on @jrlangedijk.

Dr. Rachel Berry

Rachel Berry joined Citrix from Siemens PLM where she worked as a CAD-architecture engineer, specializing in CAD and PLM joint-product development with Autodesk, Siemens NX, Ansys Workbench, SolidWorks and others. Rachel holds a PhD in Electromagnetic Field Simulation and Image Processing from Cambridge University. Rachel is now Product Manager for HDX 3D Pro Graphics at Citrix. Follow Rachel on @rhbBSE or her deprecated Citrix CAD Virtualization blog: https://virtuallyvisual.wordpress.com.

Dane Young

Dane Young (@youngtech) is a Citrix Technology Professional (CTP), VMware vExpert, technology enthusiast and Virtualization Practice Manager at Entisys specializing in application, desktop, and server virtualization technologies from Citrix, Microsoft and VMware. Dane maintains a virtualization focused blog that can be found at blog.itvce.com. He holds a number of certifications along with a Master of Business Administration and Management Information Systems degrees.

Thomas Poppelgaard

Thomas Poppelgaard was awarded Microsoft Most Valuable Professional (MVP) in 2015, Citrix Technology Professional (CTP) since 2013 and RES Software Value Professional since 2013. Thomas has 19 years of experience in IT. Currently he is working as a Technology Evangelist / Independent Consultant with his own company Poppelgaard.com. His work includes advisory, business development, IT strategy, presales, architectural designing, planning and implementation solutions and troubleshooting around Citrix, VMware, Microsoft, NVIDIA, AMD and INTEL products. He is involved in Proof of Concepts to help proving why businesses should adapt new technologies. Specialized in how CAD/CAE/CAM/entertainment/media applications can be virtualized with Citrix/VMware/Microsoft (SBC/VDI technologies) with NVIDIA/AMD/INTEL. Helping accelerate business why the GPU is important for virtualization for common applications and for high end graphical applications. Thomas Poppelgaard is a regular speaker at NVIDIA GTC, BriForum, Citrix Synergy, E2EVC, Citrix User Group in England, Norway, Finland, Holland and VMUG. Check his blog www.poppelgaard.com or follow him on twitter @_POPPELGAARD.

Pat Lee

Pat Lee is the Senior Director, Mobile Experience for VMware Desktop and Application products including VMware Horizon. The Mobile Experience team is responsible for 3D graphics, remote display protocols, remote device access, desktop clients, thin clients, web clients, and mobile clients. Pat has designed and launched multiple award winning products in data protection, virtualization,

and desktop and application delivery. Pat earned a BA in Physics from the University of California, Berkeley. Follow @patlee on Twitter.

Quotes

"This paper is great because who has the time to go research all these things one-by-one, let alone actually set them up and take measurements? (Well, these guys do I guess!) I love that they go into the background of how 3D graphics are used in desktop virtualization today, like how GPUs and APUs work, how the different types of GPU virtualization work, etc. They also cover the details of how the various remoting protocols work, like client-side versus host-side rendering, GDI remoting, etc. Really anyone who's delivering remote Windows desktops needs to read this paper." **Brian Madden, speaker, author, blogger,** @brianmadden

"Virtualization industry experts Benny, Ruben and Shawn have produced a comprehensive and up-to-date document covering exactly what technologists need to know about '3D Graphics for Virtual Desktops'. It's a thorough, independent and well-researched technical deep dive into strategy, vendors, solutions, features, qualifying questions and much more. This will be an invaluable source of information for CTO's and technical architects looking to realize the many benefits of virtualizing and remoting 3D graphics applications." **Derek Thorslund, Director of Product Management, Citrix Systems,** @DerektCitrix

"Shawn, Benny, and Ruben have created something special with their 3D Graphics for Virtual Desktops Smackdown whitepaper. This document is a deep dive into the technology and product landscape that is useful to newcomers and experienced practitioners alike. And they don't disappoint with the new 3D Graphics for Virtual Desktops Smackdown. Once just the domain of a subset of power users, 3D graphics are now a requirement for many different users and use cases. Shawn, Benny, and Ruben do an excellent job reviewing the technology behind 3D graphics on virtual desktops and providing an objective view of the product offerings. It's a must-read for any CIO or CTO considering virtual desktops." **Kit Colbert, CTO VMware,** @KitColbert

"Benny, Ruben and Shawn demonstrate great breadth of knowledge in the field of virtualization as well as in-depth analysis of the underlying technologies. "Smackdown" is an essential reference to the technology enthusiast interested in virtualization and the technology manager examining virtual environments for the datacenter." **Tonny Wong – Product Manager AMD**

"Most desktop virtualization projects start by standing up solutions in the lab, deciding they're good enough, and proceeding with a deployment. Only after scaling up the number of users and use cases does anyone really start to look under the hood to see how things work. This <white paper / book> gives you everything you need to know before you start, diving into all the combinations of desktop virtualization platforms (Citrix, VMware, MainFrame2, Microsoft, etc…) and protocols, exploring the different graphics engines (GDI, GDI+, OpenGL, DirectX, Flash, etc…), explaining the differences between graphics cards (AMD, NVIDIA), and how everything (literally everything!) relates to each other. The discussion around how to assess your 3D graphics needs and performance, not to mention the comparisons of different solutions is unparalleled in the industry, and everyone that reads this can take away useful information." **Gabe Knuth, author, blogger and editor at TechTarget,** @GabeKnuth

"Shawn, Ruben, and Benny expertly paint a picture of a fast-moving landscape of remote 3D graphics. It's an instant reference guide and a must-read for anyone contemplating remote deployments of 3D graphics apps." **Nikola Bozinovic, CEO and founder Frame, http://www.fra.me**

Modern Workspace

Flexible work styles, device proliferation, bring your own, collaboration, workspace management, mobility, AnyApp, AnyCloud, containers, application and desktop delivery are the main trends in the 'Modern Workspace'. 3D graphics for virtual desktops and applications is part of the application and desktop delivery solutions stack in the modern workspace. Virtualization is incorporated in several of these trends. Virtualization is nothing more than the decoupling of IT resources. The forms of virtualization that are most frequently applied include network, storage, server, application and desktop virtualization. Application and desktop delivery is a process which has the goal of offering applications independent of location and device, so the business consumers can work onsite, online, off site, offline, anywhere, with any (own) device and at any time. The dynamic delivery of applications is an essential functionality and part of a broader strategy of the modern workspace.

*"**GPUs are prevalent in practically every single endpoint device** that business consumers use on a daily basis. However, in virtual desktop environments the presence of a GPU is currently a very small percentage of desktops. Given the improvements that a GPU provides to a virtual desktop, it's not a question of whether or not you'll use GPUs in virtual desktops but rather when you'll use them."*
Shawn Bass - EUC CTO of Desktop Technologies - VMware

Application and Desktop Delivery

One of the fundamental questions in application and desktop delivery is the following: "What is the execution platform for the applications and where is my data stored?" Within the execution platform, system resources such as the CPU, GPU, memory, storage and network are used in order to execute the windows/Linux, web-architected, rich mobile and mobile web-applications. The most frequently used execution platforms include the following: tablet, smartphone, desktop, laptop and desktop virtualization with both Virtual Desktop Infrastructure (VDI) and Session Virtualization (Server Based Computing).

The choice of an execution platform is the most fundamental decision and it defines the application and desktop delivery strategy. Applications are either executed on a local device or centrally in a private or public datacenter. Each execution platform has its own characteristics. The theories: "Less is more", "Cut out the exceptions" and "Manage diversity" should always be in mind. An execution platform is great; but if there are no applications available on the platform, the platform is of no real value to the business consumer, the end-user. The other questions which needs to be answered is: "How will the (Windows) applications get onto the execution platform?!" A number of solutions exist for making Windows, web-architected (SaaS) and mobile applications available on the platform. The most commonly options used for Windows applications include installation, application virtualization, and more recently, layering.

"Modern applications, window managers and browsers benefit from a GPU, which massively improves user experience. In the near future, 'GPU inside' will be the standard for every Virtual Desktop solution." **Dr. Benny Tritsch |** speaker, author, blogger

It is interesting to see more and more scenarios where high-end graphics and resource-intensive (Windows) applications within desktop virtualization solutions are being used. Besides of the high-end graphics a lot of today's modern Operating System and application benefit or require a GPU. Modern laptops, desktops and tablets do have a GPU and high speed storage - solid state technology,

why equip a modern (virtual) workspace with less? The goal of this document is to explain the use-cases, business benefits, various solutions and differences between the solutions.

Terminology

Before doing a 'deep dive' into 3D graphics for virtual desktops, we think it's important to have an overview of all the 'Application and Desktop Delivery' solutions. In the following we will give you a complete outline of the existing solutions. This chapter, however, is not aimed at describing all application scenarios or their technical advantages and disadvantages, but giving a general idea of the common terminology and the state of the union in the application and desktop delivery segment, independent of vendors.

Secure Access

Secure Access solutions provide (untrusted) devices secure access to corporate IT resources. A Secure Access solution could be a full (SSL) VPN solution or a Gateway Services which is targeted for Server Hosted Desktops. Solutions that can be used to realize secure access scenarios include Cisco ISE, Citrix NetScaler Gateway, Juniper SSL VPN, Microsoft Remote Desktop Services Gateway, Microsoft UAG and VMware Access Point Security appliance.

Mobile Application Delivery

Rich Mobile applications running natively on Apple iOS, Google Android, Blackberry or Windows Phone and Modern Apps on e.g. Windows 10 are delivered by the Mobile Application delivery solution. In enterprise customer scenarios this function is incorporated in most of the Enterprise Mobility Management solutions but it can be a more consumer focused application store as well. The application store is the interface for application access, rich mobile application delivery, self-service and usage reporting functionality.

Web Application Acceleration

Web Application Acceleration appliances or application delivery controllers accelerate and secure web-architected applications. All of us are encountering these solutions every day. Large public facing internet applications, such as Amazon and eBay, all make use of these devices. Web Application Acceleration solutions are not just useful for large public organizations; you can also use them for your own web applications. Solutions that facilitate web application acceleration and security include products such as Citrix NetScaler and F5 BigIP. Functions of Web Application acceleration are:

- Secure access to SaaS and Web resources, represented by the shield symbol
- Fast and optimized access to web applications, presented by the accelerator symbol

Connection Broker

A connection broker determines which hosted remote applications and desktop will be available to a user. When using a hosted virtual desktop infrastructure for this, it is possible to either designate dedicated desktops or a pool of remote desktops. The desktop broker can automatically create, remove or pause remote desktops.

Application Virtualization

Application virtualization can make applications available to desktops, laptops, server-hosted VDI and Remote Desktop Session Host (TS-Terminal Server) platforms. The applications are executed and often isolated on the target platform, without needing to make any persistent modifications to the platform. The advantages of application virtualization include: installation, upgrade, roll-back, delivery speed and the ease of application support and management. The physical installation of applications is no longer necessary, eliminating the possibility of conflicts. The result is a dynamic application delivery infrastructure.

Application Layering

Application layering has recently emerged as an alternative way to package and deliver applications separate from the operating system, without having to reinstall them on every desktop. Applications can be layered by running a standard setup procedure. The changes in files, directories, and registry keys made by the installation procedure are captured as a "layer." When the layer is assigned to one or more desktops, its files and registry keys are overlaid (layered) with the Windows operating system layer and all other application layers forming one transparent Windows desktop to the user.

Layered applications appear in Windows Add/Remove programs and they look as if they are natively installed in the Windows desktop. The advantages of application layering include simple installation, upgrade, and roll-back; ease and speed of application packaging. In addition, this technology could provide support for Boot 0 applications, device drivers, and other complex applications that cannot be virtualized with traditional application virtualization; and interoperability with all other applications and Windows itself. A key advantage of software layering versus application virtualization is typically a higher rate of application compatibility and increased flexibility. The isolation introduced by Application Virtualization is both a blessing (for the apps that need it) and a curse (for those that don't).

OS Provisioning

OS Provisioning, or Machine Based Imaging, allows workstations to boot up and run from a central image. A single image can be used simultaneously by multiple workstations. The advantage of this is that complete operating systems, including applications and agents, can be made available quickly and securely. It is possible to make a single image available to multiple VDIs, RDSH, Client Side Virtualization and physical desktop environments without causing conflicts. As a result, it is possible to upgrade or roll-back an OS quickly, simply, and without significant risks. When virtual desktops use OS streaming, (valuable) storage is saved, storage performance is offloaded and the management of virtual desktops becomes relatively simple. This means that virtual or physical machines using OS Provisioning can become stateless devices.

Server-Hosted Desktop Virtualization (VDI)

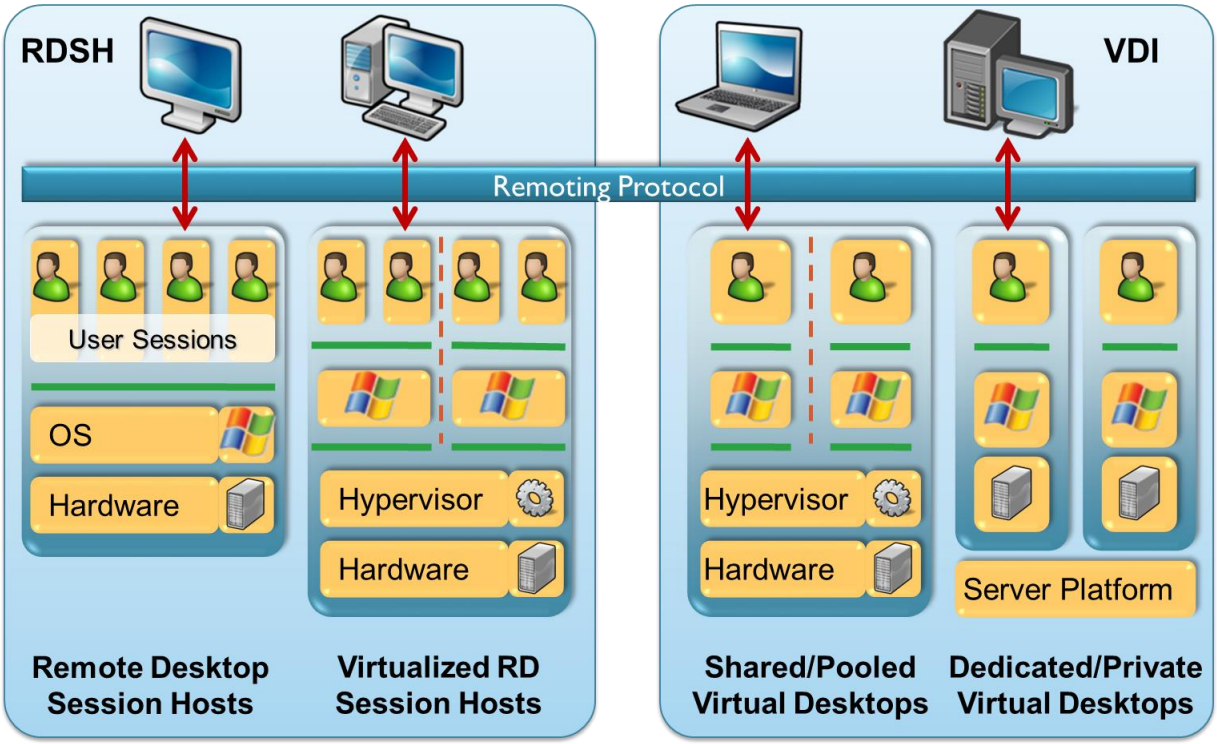
3D graphics for virtual desktops can be a server-hosted VDI solution. It provides each (virtual) machine with (GPU) graphic performance to run multimedia, 2D/3D, office, browsers, design and engineering applications. The GPU can be shared, dedicated or virtualized for the Virtual Machine or Terminal Server environment. Display data is presented to the client device via an optimized remote display protocol. To ensure that the end-users experience the best possible performance, the bandwidth, latency, or local (software/hardware) components have to meet extra requirements.

Session Virtualization (RDSH)

Session Virtualization, also known as Terminal Services or Remote Desktop Session Host (RDSH), is a solution for the remote access to desktops and applications that are run on a terminal server in a data center. With this virtualization variant, every user has his or her unique (terminal server) session, but not his or her own virtual machine. Access to the desktop or application is not tied to a location or end-user machine, and programs are executed centrally on the terminal server. The data appears on the client screen through a remote display protocol such as Microsoft RDP/RemoteFX, Citrix ICA/HDX, VMware PCoIP/Blast.

Remote Desktop Services consists of various infrastructure components for management, load balancing, session control and support. It has the advantage that applications are made available quickly and securely, the TCO is low, and applications can be accessed irrespective of location or work place.

While this document is not the proper place for a discussion of the pros and cons of RDSH vs VDI, it is important to note that RDSH does have a greater chance of limitations around application compatibility due to being based on a server operating system, whereas VDI is most often delivered via a client operating system.



Client-Side Desktop Virtualization

Client-side desktop (CSV) virtualization is a solution where the virtual machines run locally on the client endpoint device. The client hypervisor ensures that each virtual machine is hardware-independent, and makes it possible to simultaneously use a number of virtual machines at the same workstation. The hypervisor plays an essential part in client-side VDI solutions while the management portion handles synchronization, policy, enforcement and management insights.

The two types of Client-side Desktop Virtualization solution are:

- Client hosted hypervisor is installed and runs as an application on the operating system (be that Windows, Mac OS X or Linux) of the end device. This offers great flexibility of endpoint hardware and operating system compatibility at the expense of less performance.
- 'Bare-metal' client hypervisor acts as the device's base operating system and must be installed before other operating systems. This offers great performance at the expense of more limited hardware and operating system compatibility. Since a bare metal solution requires a low level installation on the target device, it is rarely an acceptable solution for the modern BYO (Bring Your Own) model of compute.

The most important differences between the two types of client-side desktop virtualization solutions are around usage in BYO scenarios, hardware support, performance, manageability and end-user experience.

Workspace Aggregation

The term Workspace Aggregator is used to describe software that unifies the delivery of multiple application or desktop types such as:

- Native mobile applications
- Software as a Service (SaaS) web applications
- Windows applications delivered through application virtualization or direct installation
- Local desktops, server-hosted virtual desktops (VDI) or published desktops in RDSH/Terminal Services

A workspace aggregator ideally evaluates the user's device to determine which applications are available for this user on this particular device and at this particular time (context-aware access). In addition to application delivery, workspace aggregators provide secure file system (data) access from a broad range of devices. A workspace aggregator simplifies the deployment and life cycle management of applications. As the single point of access, (de)provisioning, auditing and monitoring is easily accomplished through security rules and policy enforcement. Workspace aggregators that are available at the market today also Identity Management, Self Service and provide Single Sign-On (SSO) capabilities to applications.

Client management

The majority of professional IT organization use a client management solution, as it is needed to facilitate things such as OS deployment, patch management, application and client deployment, asset management, service desk integration, and remote control.

Functions of Client Management (in the context of Application and Desktop Delivery) are:

- Deliver and manage Windows/Linux applications
- Deliver and manage Operating systems, e.g. Linux, Windows Client, Windows Server

Vendor Matrix – Who Delivers What?

Vendor	Product	Solution
Citrix	XenDesktop	VDI
Dell	vWorkspace	VDI
Ericom	PowerTerm WebConnect	VDI
Microsoft	VDI with RemoteFX	VDI
NICE	DCV	VDI
RedHat	Enterprise Desktop Virtualization	VDI
NimBoxx	VERDE	VDI
VMware	Horizon (formely Horizon View)	VDI
Citrix	ICA/HDX	Remote Display Protocol
Dell	EOP/RDP	Remote Display Protocol
Ericom	RDP/Blaze	Remote Display Protocol
HP	Remote Graphics Solution (RGS)	Remote Display Protocol
Microsoft	RemoteFX/RDP	Remote Display Protocol
NVIDIA	GRID	Remote Display Protocol
Oracle	ALP	Remote Display Protocol
RealVNC	RFB	Remote Display Protocol
Nimboxx	SPICE	Remote Display Protocol
VMware	PCoIP/Blast	Remote Display Protocol
Teradici	PCoIP	Remote Display Protocol
Citrix	XenDesktop HDX3D Pro	Server Hosted- Physical Desktops
Ericom	PowerTerm WebConnect	Server Hosted - Physical Desktops
HP	Remote Graphics Software (RGS)	Server Hosted - Physical Desktops
VMware	Horizon View	Server Hosted - Physical Desktops
Teradici	PCoIP	Server Hosted - Physical Desktops
Citrix	XenClient (EOL Sept	Client Side Virtualization, bare metal - centrally managed
Microsoft	Windows 8/10 Client Hyper-V	Client Side Virtualization , bare metal - not managed
Nimboxx	LEAF	Client Side Virtualization , bare metal - centrally managed
Bromium	vSentry	CSV, Security
Citrix	DesktopPlayer for Mac/Windows	CSV, Type-2 hypervisor - centrally managed
Intel	DeepSafe	CSV, Security
VMware	Player/Fusion/Workstation/Flex	CSV, Type-2 Hypervisor – Flex centrally managed, the rest not managed.
Citrix	Provisioning Services / MCS	OS Provisioning
Citrix	Personal vDisk / AppDisk	OS Provisioning++
Double Take	Flex	OS Provisioning
Dell	Streaming Manager (WSM)	OS Provisioning
Dell	HyperDeploy	OS Provisioning
VMware	Horizon View Composer	OS Provisioning
VMware	Horizon Mirage	OS Provisioning++
Unidesk	Unidesk	OS Provisioning++

Nimboxx	Dynamic Gold Imaging	OS Provisioning
Microsoft	Windows Server 2008R2/2012/2012R2	Session Virtualization
Citrix	XenApp	Session Virtualization++
Dell/Quest	vWorkspace	Session Virtualization++
VMware	Horizon View	Session Virtualization++
Cisco	ISE	Secure Access
Citrix	NetScaler Gateway	Secure Access
Ericom	Ericom Secure Gateway	Secure Access
Juniper	SA / MAG	Secure Access
Microsoft	Unified Access Gateway / RDG	Secure Access
VMware	View Security Server, Access Point	Secure Access
Cameyo		Application virtualization
Numecent	Application Jukebox	Application virtualization
Microsoft	App-V	Application virtualization
Symantec	Workspace Virtualization	Application virtualization
Spoon	Spoon	Application virtualization
VMware	ThinApp	Application virtualization
Citrix	AppDisk/Personal vDisk	Application layering
FSLogix	Apps	Not layering, but achieves similar goals
LiquidwareLabs	FlexApp	Application layering
VMware	Mirage	Image and Application management
VMware	AppVolumes	Application layering
Unidesk	Unidesk	OS delivery and application layering
IBM	BigFix	Client Management
LANDesk	Client Management Suite	Client Management
Microsoft	System Center Config Manager	Client Management
Novell	ZenWorks Configuration Mgr	Client Management
RES	Automation Manager	RunBook Automation
Symantec	Client Management Suite	Client Management
Citrix	Storefront	Workspace Aggregator
Centrix	Workspace Universal	Workspace Aggregator
VMware	Horizon Workspace / IDM	Workspace Aggregator
AMD	FirePro	GPU enabling technology for 3D Graphics DV
Intel	Intel Graphics Virtualization Technology (Intel GVT) / IrisPro	GPU enabling technology for 3D Graphics DV
Intel	Intel VT for Directed I/O (Intel VT-d)	Enabling technology for direct assignment of virtual machines for virtual functions
HP	Remote Graphics Software	Enabling technology for 3D Graphics DV
NVIDIA	GRID, Quadro and Tesla	GPU enabling technology for 3D Graphics DV
Teradici	PCoIP	Enabling technology for 3D Graphics DV
Amazon	AWS (G2 instances)	Enabling technology for 3D Graphics DV - DaaS
Fra.me		3D Graphics DaaS
Citrix	Workspace Cloud	3D Graphics DaaS (basic 3D graphics)
OTOY	Cloud Workstation	3D Graphics DaaS
VMware	Horizon View Air	3D Graphics DaaS (basic 3D graphics)

Why VDI – Desktop Virtualization 101

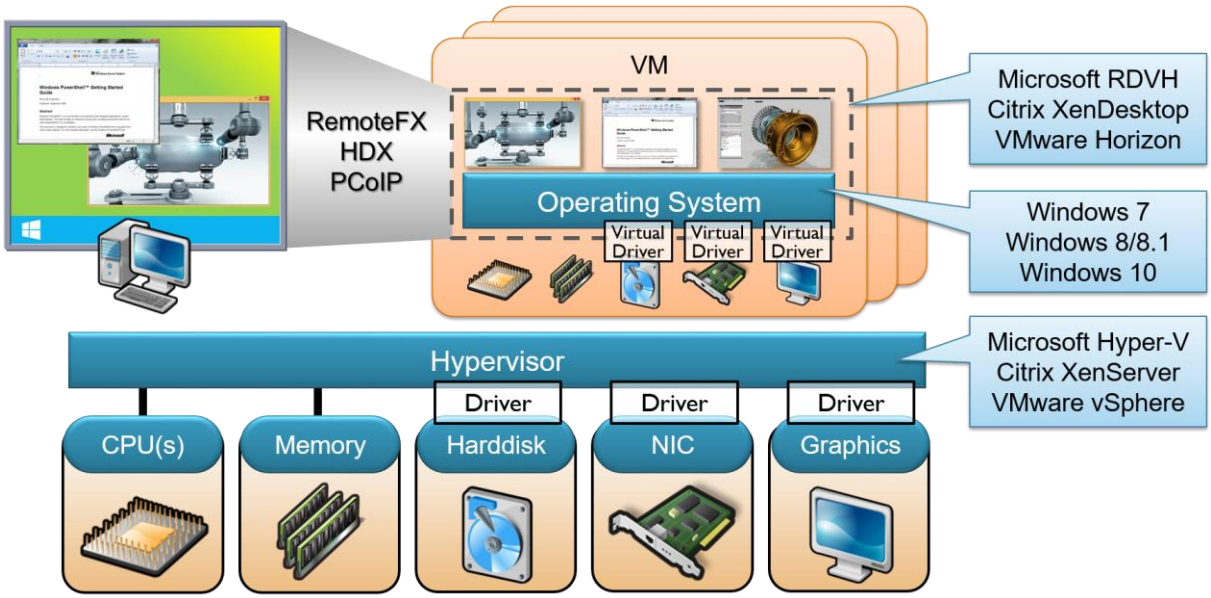
The previous chapter gave an overview of 'Application and Desktop Delivery' solutions. This chapter describes 'Desktop Virtualization' in more detail.

Server Hosted Desktop Virtualization Directions

The different server hosted desktop virtualization solutions are outlined in the following paragraphs.

Server Hosted Virtual Desktop (VDI)

Server Hosted Virtual Desktops (SH-VDI) is a solution for accessing Windows desktops that are executed remotely on a virtual machine in the datacenter. The Virtual Infrastructure ensures availability and manageability. Other frequently used terms for this type of desktop virtualization include: Virtual Desktop Infrastructure (VDI), Remote Desktop Services Virtualization Host.



Non-Persistent, Persistent and Layered Virtual Desktops

Typically, server-hosted virtual desktops have come in two different variants: non-persistent (pooled) desktops, in which users are given a fresh / new virtual machine every time they login, and persistent (private) desktops, in which users are assigned a dedicated virtual machines that remains the same every time they login. Non-persistent (pooled) desktops are often called stateless desktops because they will always revert back to their original state.

Persistent (personal) desktops are often called stateful desktops because they give users the freedom to install software (rights permitting) and make workspace-related adjustments by saving changes i.e. the state of the workspace, between sessions. Keep in mind that there are different levels of user personalization. Some desktop items like icons and wallpaper can be made persistent with a user profile or a workspace management product. You will not need to implement stateful desktops to achieve basic customization of the virtual desktop. However, if you want the user to be able to make deeper changes to the desktop (like having different installed desktop applications) then this is a use case for persistent VDI. As an alternative, a layering solution can be established to accomplish the same goals.

Stateless (Pooled) Desktops

The advantages of non-persistent (stateless) desktops are:

- Very simple to roll out and fairly easy to update the base images (think applying monthly OS hotfixes)
- The virtual desktops are guaranteed to be 100% identical
- The user always has a clean desktop (no registry clutter over time)
- Less effort is required by the management team due to the standardization of images
- Less storage is required because a single base operating system image can be shared across many desktops
- Image management may require a new method of maintenance vs existing PCLM tools

The major disadvantage of non-persistent desktops is that any customizations made by users and any applications that are delivered outside of the base image are lost after each desktop reboot. In an effort to make stateless, non-persistent desktops act more like persistent desktops, VDI vendors have recommended profile management tools to restore user customizations each time users log in to their new virtual machines, and application virtualization tools to deliver different applications outside of the base gold image. User-installed applications and deep OS and application security settings that live outside of profiles cannot be captured and restored by profile management tools.

Stateful (Personal) Desktops

The advantages of persistent (stateful) desktops are:

- Freedom to install software within the desktop (rights permitting)
- Maintaining all changes to the operating system between reboots
- Maintain desktops using the exact same tools you manage your existing desktops

Traditional Client Management solutions can be used to deploy, maintain and support personal desktops the same way as laptop, Ultrabook and desktop scenarios.

The major disadvantage of persistent desktops is the higher (storage) cost. Most enterprises cannot afford the high costs of storage required to implement full-sized, thick-provisioned virtual machines for every user. Also, since every desktop is unique, this can introduce additional unique troubleshooting scenarios for applications, etc.

Layered Desktops

Another technology has recently emerged, that gives administrators a new virtual desktop provisioning and management option for VDI: “layering.” Layering combines the benefits of stateful and stateless desktops in one solution. With layering, a non-persistent virtual machine is assigned to each user, guaranteeing that systems are consistent between reboots and allow IT a better way of managing image updates. Yet, unique departmental applications, user installed applications and all customizations will be remembered through reboots and base image updates, using dedicated, injected layers. This makes the layered virtual desktops stateful, while still providing common base image updates shared across multiple systems.

The advantages of layered desktops are:

- Freedom to install software within the desktop (rights-permitting and IT policy approved)
- Maintaining all changes to the operating system between reboots (controlled via IT policy)

- The simple roll-out and update of base images
- A virtual desktop is guaranteed to be 100% identical at the OS and application levels (in the base image layer). Individual departmental app layers and/or user layers can allow user/application deviation as needed
- The user can always be reverted back to a clean desktop (controlled via IT policy)
- Less effort is required by the management team due to the standardization of images, simpler application packaging and ability to rollback or "undo" OS and application patches
- Less storage capacity is required because a single base operating system image and single image of common applications can be shared across many desktops

A disadvantage of layering is that some of the vendors' solutions are supported in server-hosted VDI environments only. However, the fact that there are now hundreds of real world customer implementations of layered desktops, many with over 1,000 users in production, indicates that VDI-only layering solutions are being generally accepted by the mid-market.

VDcry – the Challenges

Reading the 'Why VDI!?' paragraph provided you with a general overview and better understanding of the benefits of VDI. Every solution has its weaknesses or challenges, as does VDI. Documenting challenges doesn't mean it cannot be done, it just means that attention must be paid to ensure the solution outcome is positive. The attention points for VDI are:

- End-user experience (UX) and perceived performance of the solution
    - Voice, webcam, VOIP, conferencing solutions usability and supportability
    - The requirement for networking; bandwidth, QoS, jitter, latency and packet loss
    - Usability of rich media; audio/video, Adobe Flash and H.264 content
    - 2D/3D graphics in LAN and WAN
    - Delivery of hosted applications and desktops to mobile and touch/tablet oriented devices
    - Peripheral support for USB – 3DSpacemouse, Wacom tablets, gaming controllers, haptics
- The overall cost reduction of the technology stack
- The overall complexity in the complete technology stack
- Streamline system management between VDI and classic clients
- Ease of migration to the technologies and initial start-up costs
- Vendor lock-in and Single-vendor stack considerations
- Monitoring the entire infrastructure end-to-end
- Legacy investments in hardware and opportunities to repurpose, budget cycles for hardware refresh
- Total Cost of Ownership and Return On Investment
- Licensing (SA, VDA, SPLA). These are especially important in Bring Your Own scenarios because Microsoft licensing is a major challenge. Service Provider License Agreement (SPLA) for VDI, I a huge challenge today
- The end user's user acceptance criteria. When does the solution meet the requirements for the user? Where are the requirements and who has contributed creating the requirements? If the solution does not satisfy the user, it cannot be successful. The business consumer wants a great user experience!

- Does the IT team possess the skills and mindset to succeed in the project?
- VDI has a serious impact on storage. When you don't know the impact VDI has on storage, seek help before launching the project. IOPS, throughput, sub-millisecond latency and costs are only the beginning

So in essence, user experience, 3D graphics, storage, costs and Microsoft licensing are the main barriers preventing widespread adoption of VDI. The first four of these challenges can be solved by setting the right expectations and using the proper solution.

Strategies for Using 3D Graphics in VDI

The concept of Virtual Desktops is a key component in an optimized desktop strategy. The transition to a dynamic and optimized desktop is causing many IT organizations to reevaluate traditional IT operations, deployment, delivery, packaging, support and management methods.

It is important to have a vision and a strategy around the 'Modern Workspace'. We see many organizations primarily focusing on products and vendors, but lacking a clear vision and strategy. This approach is fine for point solutions, but a proper vision and strategy is crucial for a next generation optimized desktop platform. How can vision and strategy be successful? The answer is a simple formula: Success = Vision + Execution + Adoption.

The following discussions and corresponding topics should be part of the virtual desktop strategy. This includes 3D graphics for virtual desktops.

Generic:

- What are the use cases for Virtual Desktops? Does the use case require Desktop Virtualization?
- VD-why? What are you trying to achieve? Is this a business enabler? Does it reduce the overall cost of ownership (TCO)? Will it allow your business to be more agile in how you share information and collaborate on activities? Are you doing technology for technology's sake?
- What is the business-case? What do you expect as a ROI? This can be expressed in money saved, money not spent, or even time saved (which ultimately means a boost in productivity and/or a reduction of headcount).
- Are you investigating a tactical (point) solution or a strategic solution? What is your roadmap and your timeframe?
- Is work shifting a key driver for the optimized virtual desktop? How are the roaming/flexible and mobile users within the organization facilitated? How do you take care of application and desktop delivery when the user has different access scenarios? Are you a company that insists your users aren't allowed to use Drobox, at al or are you actively seeking ways to enable your employees to work the way they want to work?
- How do you deliver applications to users in a Bring Your Own (BYO) or Choose Your Own (CYO) scenario?

Desktop strategy:

- What's your desktop delivery and migration strategy for Windows 7/Windows 10? Beyond staying secure, what is the business value to your organization in moving to Windows 10?

- What is your application and desktop delivery model? Is delivery of applications/desktop focused on SaaS, Enterprise, SMB or the consumer space? It's not uncommon for this to be a blend of technologies and that's perfectly alright.
- Is the VDI deployment targeted at a small, medium or enterprise environment? Is the solution easy to deploy and easy to maintain? Don't be afraid to look at VDI as a solution for a silo of the business if it provides the ROI/TCO to satisfy your needs. Being strategic in the beginning doesn't always mean it is the best path to proving the technology at your organization. Some companies need small wins to become motivated to be strategic about things. (i.e. don't bite off more than you can chew…)
- Do you want to deliver windows and web applications to mobile devices such as tablets and smartphones via VDI? Is this what the user wants? Try spending a day in their shoes and see if you are satisfied with that solution. Do you have an enterprise mobility strategy?
- How does the desktop virtualization solution fit into existing deployment and management tools? Will you need to retrain your software packaging / deployment team? What about your desktop / application support teams? Finally, what impact does your decision have on desktop and application delivery, and what is the influence on your high availability and disaster recovery needs?
- Is separation of operating system, applications and user preferences inside and outside the virtual desktop part of the overall strategy?

User Experience:

- What is the user experience regarding multimedia, video/audio and 2D/3D applications? Always try to look into the future with regards to this. There's nothing worse than deciding on a virtual desktop architecture and then having to re-architect it six months after it's rolled out because you were shortsighted.
- Is 3D graphics needed? Which applications benefit from GPUs? What is the perceived end user experience expectation?
- Is Unified Communications and VoIP functionality within VDI needed? Is it supported by the VDI and UC-vendor? Does the (thin) client vendor supports this as well/ don't forget how your mobility strategy is impacted by this. Also, UC/VoIP is very different inside the firewall versus outside the firewall. Do you have a list of tested VoIP/UC peripherals, such as supported softphone headsets and speakerphones? Is your UC/VoIP solution spanning long distances over poor WAN links? Have you tested these scenarios?
- What are the user expectations of the virtual desktop? Are users involved in a proof of concept and pilot? What are their acceptation criteria?
- Never conduct a proof of concept or pilot with the easiest use cases and users in your environment. It will create a false sense of accomplishment within the project and may very well land you in hot water when you are asked to scale up to thousands of users across multiple regions with very complex requirements. Always select an unbiased mixture of users from different departments with different needs to ensure you have handled challenges from every corner of the business.
- What endpoints do you support and facilitate, and what is the role of these devices for the end-user experience? Are the endpoints managed by the IT organization? It is very important to have endpoints and peripheral devices that your users are using in the environment so IT can see how it works. Even if you claim you don't support third party hardware in the

environment, you always end up supporting more than you would like to, depending upon how important the individual user is.

Infrastructure:

- Are there established concepts for secure access and secure networking? How do users connect to the virtual desktop when using a variety of endpoints (rich, thin or zero clients and mobile devices)?
- What is the impact of Secure Access and Secure Networking solutions on mobile devices while connecting to the virtual desktop? What is the user experience with these secure access solutions?
- What is the performance and bandwidth impact on the network infrastructure; LAN, WAN, WLAN and mobile networks?
- Do you use Build Your Own, Converged Infrastructure or Hyper Converged Infrastructure solutions to host the various workloads?

VDI Solution:

- Does the VDI solution needs a client/agent component on the endpoint? Is there a supported agent for the OS/endpoint? What is the User Experience with this agent? What is the feature and future roadmap of these agents? Is agentless via HTML(5) included and important?
- Is image deployment and management part of the (virtual) Desktop Strategy?
- Do you need a virtual machine-based image management solution?
- How do you design, build and maintain the (golden) Image(s)?
- Does the end-user need the ability to install and update applications? Is a user-installed applications functionality needed? Does the user have the correct privileges to install or update software?
- How are Windows applications delivered within the VDI environment? Is unattended installation, manual installation or application virtualization the selected methods, or are the applications part of the (golden) image? What is the strategy?
- Do you focus on stateless (non-persistent) and/or stateful (hybrid/persistent) images? What is, for example, the impact on storage, manageability, security, legal and business-case?
- Are you using Linux or Windows as guest OS?
- Is Windows 10 or Windows 7 your core guest OS platform? Is it x64 or x86?
- How does the solution scale? What do we need from a scalability point of view?
- Is there a validated design of the target environment?
- How do you size the virtual desktops and the corresponding infrastructure? What are the best-practices for optimizing the virtual desktop?
- What is your site topology? Is it a multi-site environment with multiple datacenters?
- Is the VDI solution as a whole highly available? Is that built-in or are additional planning and solutions needed for high availability? Are highly available virtual desktops needed?
- Is the VDI vendor a financially healthy organization? Is this important in evaluation of the vendor and the solution?
- Is there a healthy ecosystem with partners, consultancy, training and education around the VDI solution? Is this important for you?
- What is your public cloud computing strategy and how does VDI/DaaS fits in that strategy? What do you want to solve?

- Is the solution reliant on public cloud infrastructure (IaaS, e.g. Amazon EC2, Windows Azure, VMware Air)? Where is the data stored? Is US Safe Harbor a legal challenge? What is the exit strategy for your public cloud offering?

User Environment Management:

- Do you need context awareness? Based on user/role, device, location and various settings is access to application resources controlled and enforced when needed?
- How do you design and build the user's profile and his 'workspace'? Does application virtualization fit into this strategy? The 'User Environment Management' Smackdown whitepaper can be helpful.
- Do you want to integrate and run local applications in the centralized desktop environment and present centralized and local applications in one single interface to the end-users?
- What is the impact of client management and user environment management solutions in a stateless (pooled) VDI scenario? Is it supported?

Application Readiness:

- What is your application readiness assessment strategy? Are Windows 7, Windows 8, Windows 10, VDI, application virtualization and x64 included?
- What is the (business) applications vendor support policy for virtualized desktops?
- Do you have insights which of your applications are graphics and resource intensive? If not, you may need a monitoring solution.

Storage:

- What is the performance and storage impact of application virtualization in combination with VDI? Is this important from a business-case or technology perspective?
- Do you use classic storage appliances, all flash storage appliance, software defined storage or hyper converged storage? What storage optimization (IOPS/latency) technology is being used? Do you understand the benefits of Software Defined Storage (SDS) and Hyper Converged Infrastructure solutions?
- What are the storage licensing concepts for VDI solution, guest and client operating system, client access and (business) applications?
- Do we need to backup (and restore) the vDesktops?
- Is Antivirus needed? Inside the VM or as a layer on the Hypervisor? What is the real performance impact of Antivirus protection?

IT Organization:

- Is the IT department able to adopt the technology with right knowledge and skills? What subject matter experts are needed to get and keep the VDI solution up and running in production environment? Is this expertise available? Who has overview of the complete VDI solution stack?

Do you have other key questions we should address? Please let us know: team@teamrge.com

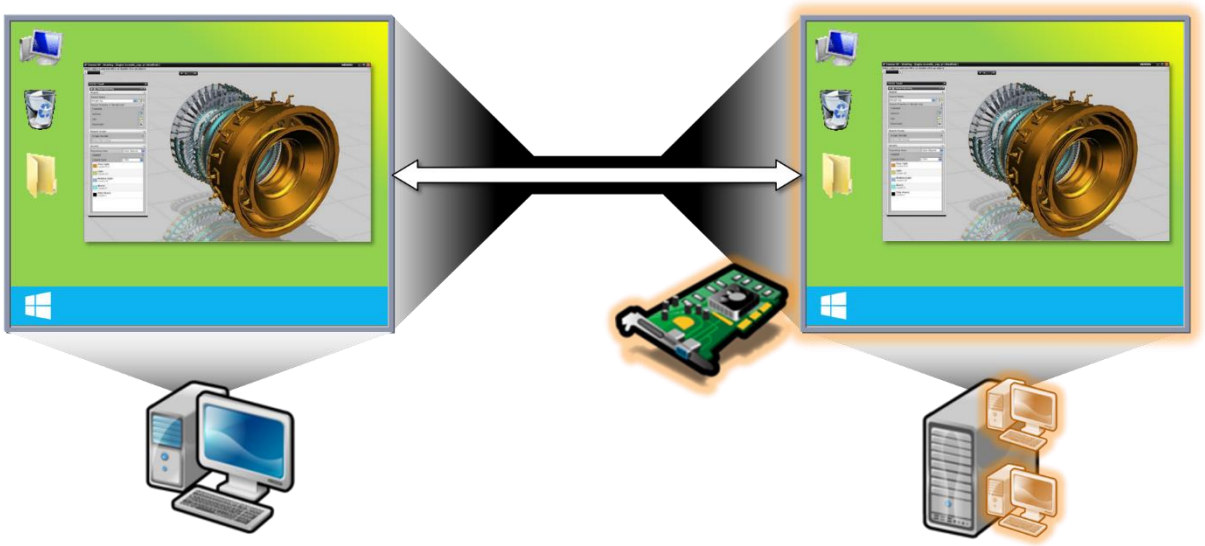
3D Graphics for Virtual Desktops

*"**User Experience is king**. GPU capabilities and blazing fast solid state storage is the standard for a modern virtual workspace!"* **Ruben Spruijt**, CTO Atlantis Computing | MVP, CTP, vExpert

Delivering 3D graphics for virtual desktops is not a niche market. Everyone can benefit from graphics with their virtual desktop. With technology evolving and increasing competition, the cost of adding 3D graphics to virtual desktops will be affordable for everyone.

Literally every PC has a GPU (Graphics Processing Unit), so why is this not the case for VDI? Most new operating systems and applications leverage GPUs while a growing number of users benefits from using high-end graphics for 2D, 3D, multimedia applications and HPC environments.

Many people traditionally associate 3D graphics with heavyweight specialist workloads such as CAD/CAE/CAM applications or rich multimedia and gaming. But in reality, 'ordinary' users (also called knowledge workers) are in fact also using high-end graphics and GPUs on physical PCs. Many standard applications such as Web browsers, YouTube players (video), Unified Communication (Lync), Microsoft Office and Windows Aero effects can benefit from GPU acceleration.



Why 3D Graphics for Virtual Desktops?

What are the business drivers for enabling 3D graphics for Virtual Desktops?

- Flexibility: Work isn't a place, it's something you do from anywhere with any device
- Access: Virtual desktop works independently of location, endpoint and network. Use the Virtual Desktop from any client endpoint; work anywhere, with any device in LAN, Wi-Fi and WAN scenarios. Support work from home, global product development teams, contractors and BYO scenarios.
- Application integration: More and more application integrates with each other, large files and datasets (sometimes Terabytes) are accessed by users all over the globe. File transfer and WAN acceleration are not the solution.
- Security and control: Information systems and data stays in the datacenter center; Protect Intellectual Property (IP).
- Freedom: Every user can have his or her own (virtual) desktop with administrative privileges when needed.

- BYO: This concept enables the delivery of applications and desktops for Bring Your Own devices or within Corporate Owned Personally Enabled (COPE) scenarios.
- High performance: High-end graphics and flash based storage solutions deliver high-end 2D/3D graphics and resource intensive applications to any device while keeping the data central.
- Reduce cost: Operating System, applications and user environment are centrally managed. Resources are shared where needed.
- Legacy: It is simple to offer legacy applications on a state-of-the-art end-user client computing platform;
- Cooling: Use thin clients with multiple screens in larger setups in the offices and keep heat generating workstations inside the cooler datacenter. Reduced noise helps to improve employee working conditions as well.
- Disaster Recovery: Multi-site and multi-datacenter around the globe is easier (not easy) to design because apps, data and desktops are centralized.

Use Cases

There are already numerous use cases for 3D graphics for virtual desktops and the number is still growing rapidly. The most common use cases today are around delivering 3D graphics and resource intensive applications for aerospace, automotive, construction, energy, film/media, engineering, hi tech, electronics, education, industrial equipment, medical equipment, oil/gas exploration and HPC environments. Various customers around the world in different market segments are using some sort of GPU-accelerated remoting technology to meet their use cases.

Note: TeamRGE is happy to connect you with these clients when appropriate.

User Classification

Perceived performance and user experience are critical elements in a 3D graphics for virtual desktop project. It is important to understand the user population, what applications or type of applications they use and what the performance and resource impact of these applications is. It doesn't really matter how exactly you classify your users and what names you use to group them, as long as you keep it consistent.

Classifying the GPU requirements of the applications is also important. Does the application require a GPU to operate, is the application GPU intensive or is the GPU just supporting the application? Below you find the common terminology to group the different 3D graphics user types in conjunction with 3D graphics for virtual desktop concepts. The details of the different 3D graphics concepts will be explained later in this document.

Designers / Engineers - Power Users

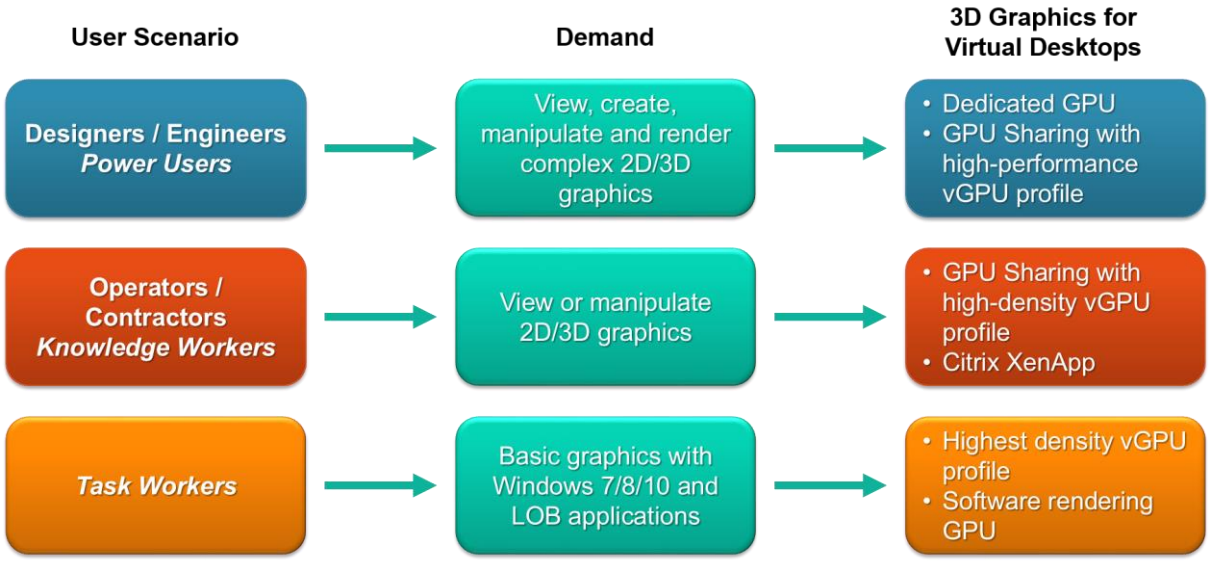
Within an organization, typically designers are the members of the most demanding user group with respect to graphics. They view, create, manipulate and render complex 2D/3D graphics. GPU acceleration for Direct3D, OpenGL, CUDA and OpenCL applications is required. Designers require high-end virtual desktops with high performance, low latency storage and dedicated GPUs for graphics acceleration. GPU pass-through can be the preferred solution for designers. Almost as demanding as designers, engineers view, create and manipulate 2D/3D graphics. Engineers require a higher-end virtual desktops with high performance storage and a high-end virtual GPU profile.

Operators / Contractors - Knowledge Workers

The users belonging to the knowledge worker category view or edit 3D models and need to access graphical applications and workflows from anywhere. This requires high-performance, solid state/flash/memory speed storage to start and use applications. Hardware accelerated graphics is needed to a certain extend. High density virtual GPU profiles and GPU sharing are suitable options for this kind of users. PLM/PDM workers are also great examples of knowledge workers.

Task Workers

The segment of users in the organization that are not engaged in professional graphics design. Hardware accelerated graphics may or may not be required to deliver business graphics, such as the Windows 10 style apps, PowerPoint transitions in Office 2013/2016, or perform light 2D and 3D work.



3D Graphics for Virtual Desktop Concepts

This chapter explains the different solutions to deliver 3D graphics for virtual desktops.

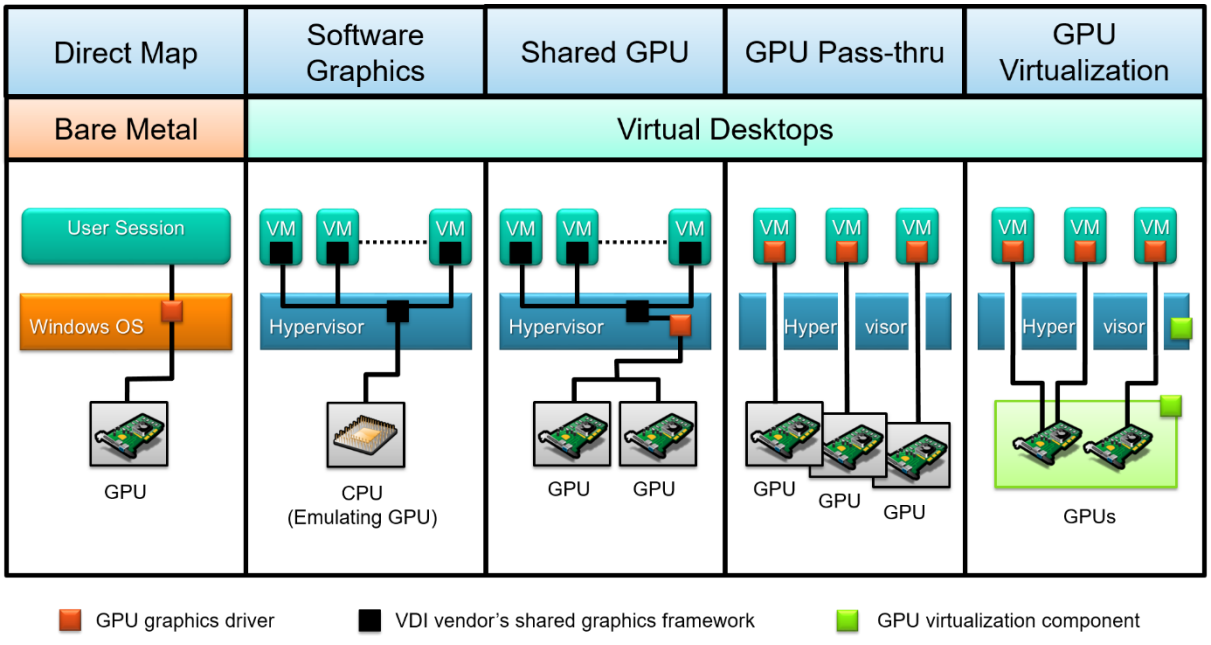
GPU and APU

Delivering high-end graphics for virtual desktops can be accelerated by using physical Graphics Processing Units (GPUs). A GPU is a highly integrated electronic circuit used in personal computers, mobile phones, embedded systems, graphics workstations, game consoles and cloud platforms. It is designed to perform data manipulation at high performance and in parallel, with the intention to create digital images in a frame buffer and send the results to one or multiple displays. In contrast to a computer's Central Processing Unit (CPU) with only few cores designed for sequential serial processing, a GPU consists of thousands of smaller cores optimized for dealing with many tasks simultaneously. In a graphics workstation or a cloud server optimized for graphics remoting, GPUs are typically present on a separate video card.

An Accelerated Processing Unit (APU, also Advanced Processing Unit) combines a CPU with additional processing capability designed to accelerate certain types of computations. Such an APU may include a GPU used for General-Purpose computing (GPGPU) and for graphics acceleration. In essence, an APU combines a CPU and a GPU on the same die, allowing for high data transfer rates between the two while keeping power consumption at a relatively low level. A good example of a state-of-the-art APU is Intel's Haswell and later CPUs with integrated Iris Pro GPU. While a dedicated GPU on a separate video card may outperform an APU in many high-end graphics scenarios, lower overall power consumption and lower price are typically advantages on the APU side.

3D Graphics for Virtual Desktop Concepts Summary

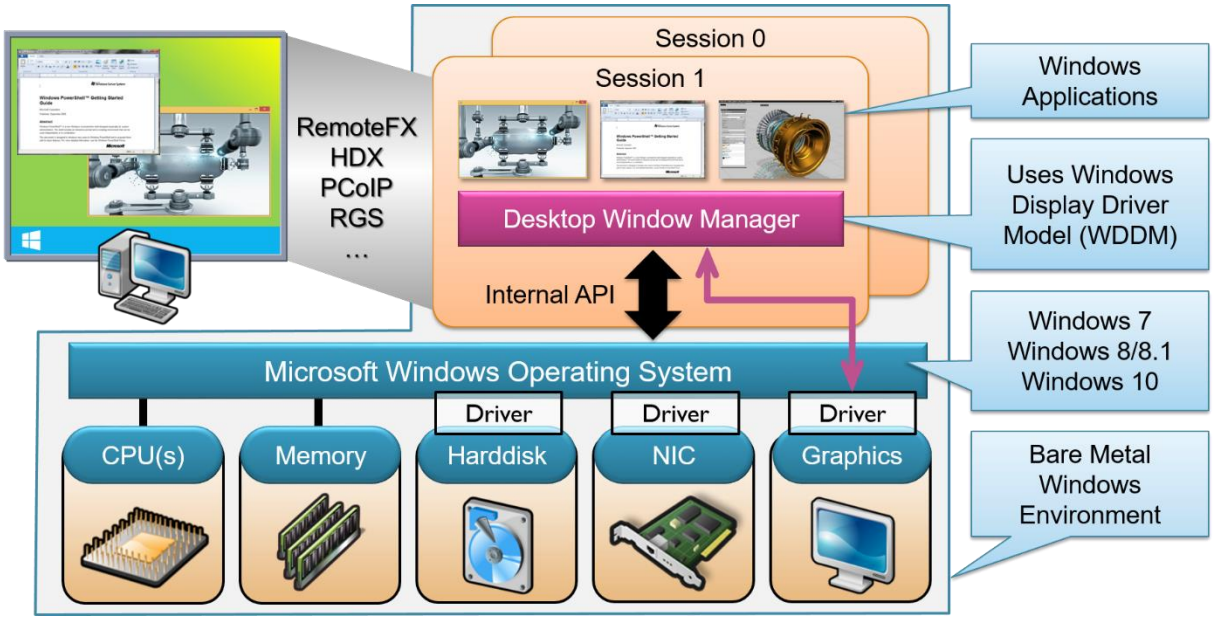
The figure below shows the five different concepts to leverage 3D graphics for virtual desktops:



The five graphics models are described in detail in the following sections.

Bare Metal Graphics Model

This model represents the (blade) PC/workstation remoting architecture where no virtualization is being used. The end user connects via a connection broker to the workstation via remote protocols such as Citrix ICA/HDX, Microsoft RDP/RemoteFX, VMware Horizon View PCoIP/Blast, HP RGS or via Teradici hardware based PCoIP. HP Moonshot, powered by Intel APU, is an example of a solution to deliver a virtual desktop running on a bare metal PC-like hardware.



Benefits:

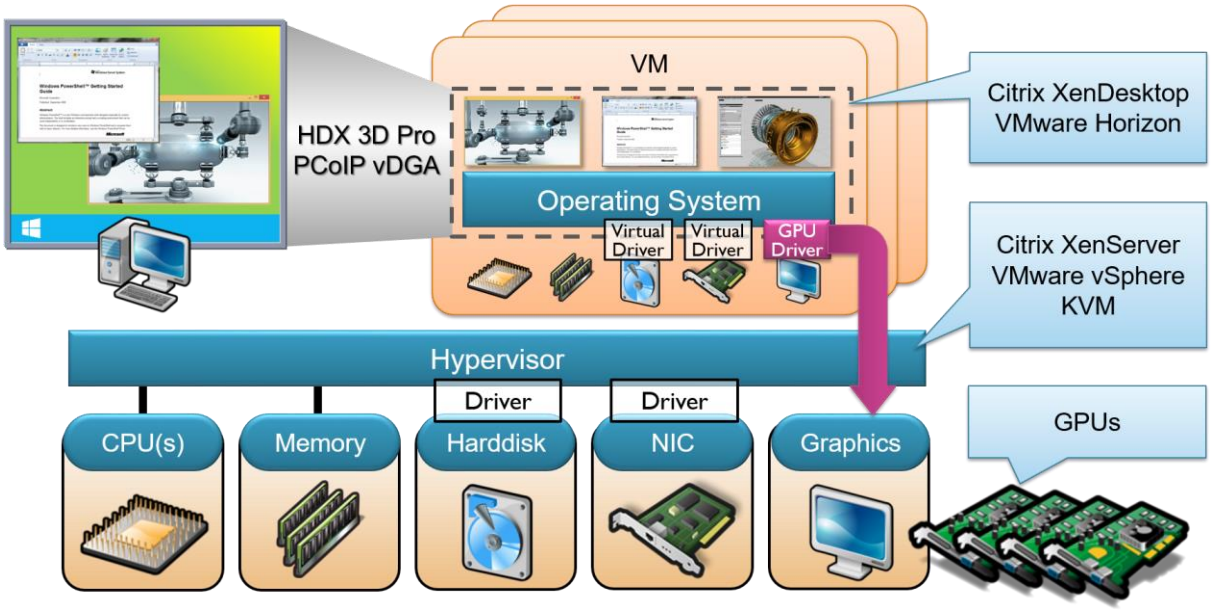
- High-end 3D Graphics and workstation class performance
- Supporting all graphics formats, such as OpenGL, OpenCL, NVIDIA CUDA and DirectX 9/10/11/12
- Full application support, native GPU drivers are being used

Challenges:

- Hardware dependent
- Less cost effective
- No Virtualization benefits such as hardware independence, resource sharing, snapshots, Live Migrations etc.

Pass-Through or Direct Attached GPU for Virtual Desktops - VDI

In this case, the virtual machine (VM) has a 1:1 relation with a GPU, meaning that the Hypervisor allows the VM full and direct access to the GPU. The native NVIDIA/AMD/Intel graphics driver is installed in the Virtual Machine. VMware call this technology “vDGA - Direct Graphics Acceleration”, Citrix calls this technology “GPU pass-through”. Independent of its name the solution offers full GPU performance and high-end graphics needed for designers and engineers.



Citrix XenServer 6.x and VMware vSphere 5.1 or higher support the pass-through or direct attached GPU for Virtual Desktops.

Benefits:

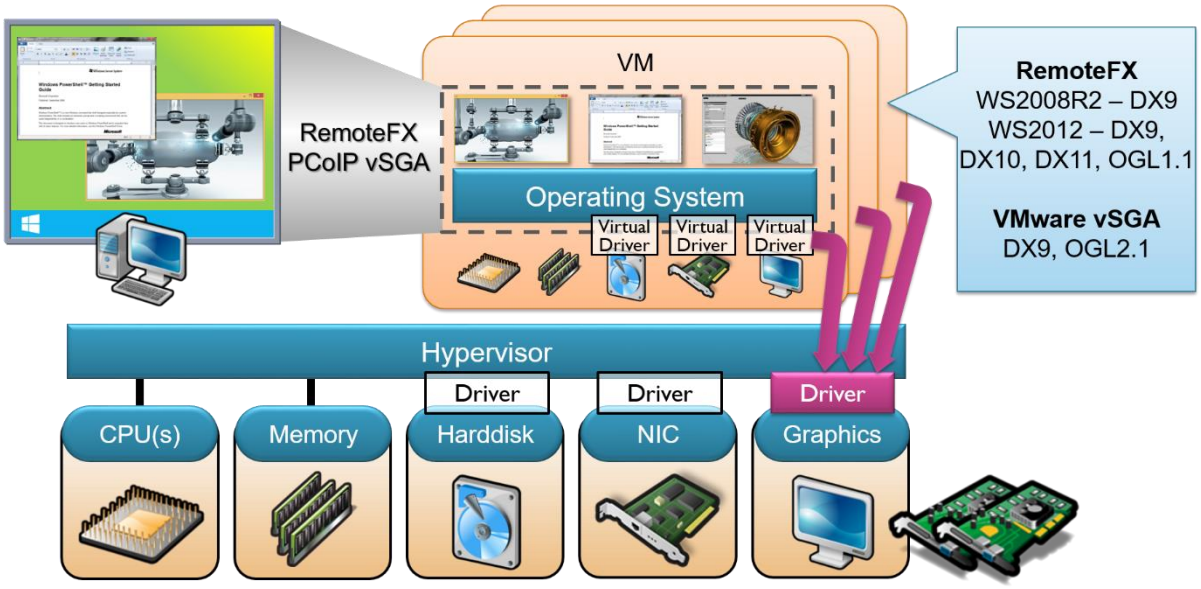
- High-end 3D Graphics and workstation class performance for virtual desktops
- Support for all technologies such as OpenGL, OpenCL, NVIDIA CUDA, DirectX 9/10/11/12
- Full application support, native GPU drivers are used
- Virtual machines with and without direct access to the GPU can be hosted on the platform

Challenges:

- Live migration of VM with pass-through devices is today technically not possible
- Hardware dependent
- Monitoring the GPU usage can only be done in guest
- Typically, the most expensive way to provide GPU-accelerated remote access

GPU Sharing for Virtual Desktops, API Intercept

GPU sharing for virtual desktops is a 3D graphics solution suitable for both VMware Horizon View (VDI) and Microsoft Remote Desktop Services. The shared GPU model is also referred to as API intercept where the GPU is managed and owned by Microsoft Hyper-V with RemoteFX or VMware vSphere with Horizon View vSGA (Shared Graphics Acceleration). The graphics API requests generated from the VMs are intercepted via the capture driver in the virtual machine. Typically, this is the synthetic graphics driver included in the integration components of the hypervisor. The graphics requests generated in the VM are redirected to the hypervisors and executed with GPU support. After the graphics processing is finished the resulting graphics data is sent back to the VM. The VM has no direct access to the GPU, as with GPU Virtualization for VDI.



The advantage of GPU sharing is its superior scalability, delivering 10s of VMs or sessions per physical GPU which allow knowledge workers and task workers hardware-accelerated access to Line-of-Business applications and Windows desktops.

Benefits:

- Cost effective solution for knowledge users and task workers who require graphics acceleration
- Scalable solution for 10s of VMs or sessions, of course depending on application and usage
- VMware vSGA implementation can switch between GPU sharing and software 3D renderer Soft3D is a VMware WDDM driver which is installed in the VM
- Allows each user to have power user performance with enhanced support for DirectX 3D and Windows Aero
- Supports live migration of virtual machines with VMware vSGA

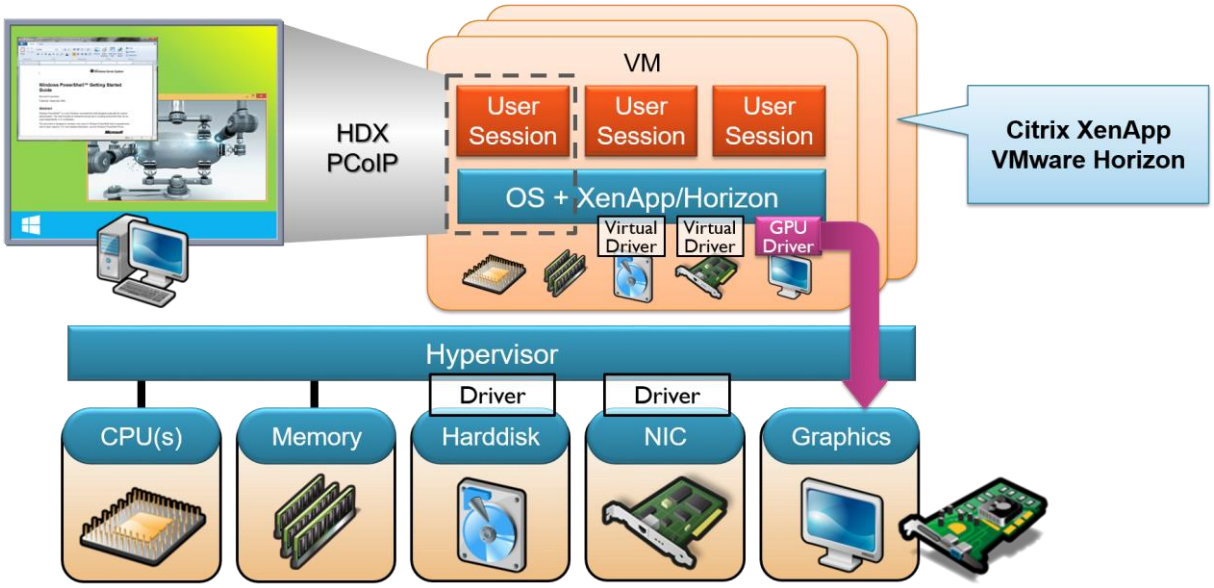
Challenges:

- The solution doesn't meet the requirements for high-end graphics where users will view, create, manipulate and render complex 2D/3D graphics
- The GPU can become a performance bottleneck as many users will share the same GPU.
- There may be a different set of supported graphics formats for multi-user server platforms (such as Citrix XenApp or Microsoft RDSH), depending on if the server is virtualized or running on bare metal.
- Application compatibility issues due to limitation of 3D APIs supported
    - Limited or missing support of certain OpenGL versions
    - In some cases, DirectX support limited to version 9

GPU Sharing for Virtual Citrix XenApp and VMware Horizon

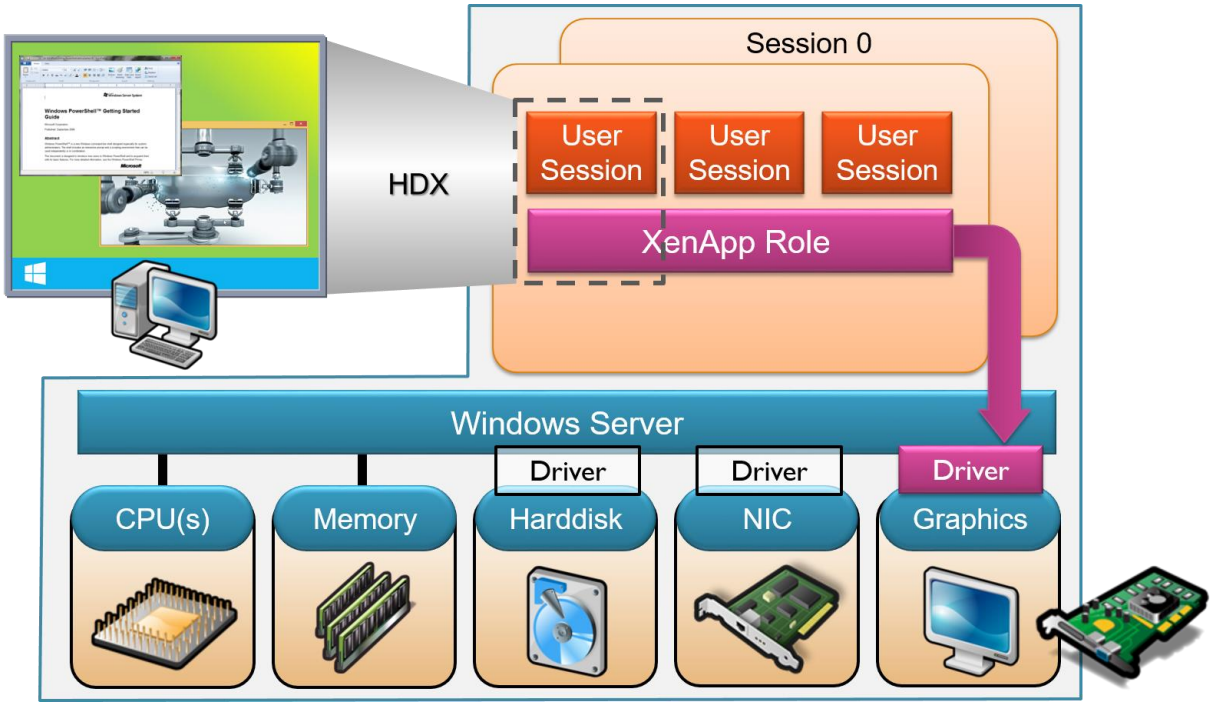
Citrix were the first who introduced a slightly different model to share GPUs on XenApp. The foundation is a multi-user Operating System such as Windows Server 2012 R2 in RDSH role deployed on a hypervisor with a physical graphics card. On such a system Citrix XenApp and a native graphics driver need to be installed. Now the users of all interactive sessions can access the physical GPU or a

virtual GPU (part of a physical GPU). Only recently, VMware released the same functionality with their Horizon product.



GPU Sharing for Physical Citrix XenApp

In this model, a multi-user Operating System such as Windows Server 2012R2 and Citrix XenApp on bare metal are the foundation. 3D applications are installed and published in a hosted shared desktop or as seamless applications. The Citrix XenApp server needs to have access to a GPU, which can be the scenario with bare-metal or a virtual machine with GPU pass-through. It is important to note here that in a pure Microsoft stack, RDSH on top of the Hyper-V hypervisor only supports bare metal GPU access but not PCI device pass-through. RDSH hosted on alternative hypervisors does provide PCI device pass-through access for VMs.



In order to leverage a hardware graphics adapter for RDSH, you must enable a GPO setting named "Use the hardware default graphics adapter for all Remote Desktop Services sessions" on Server 2012/2012 R2.

Benefits:

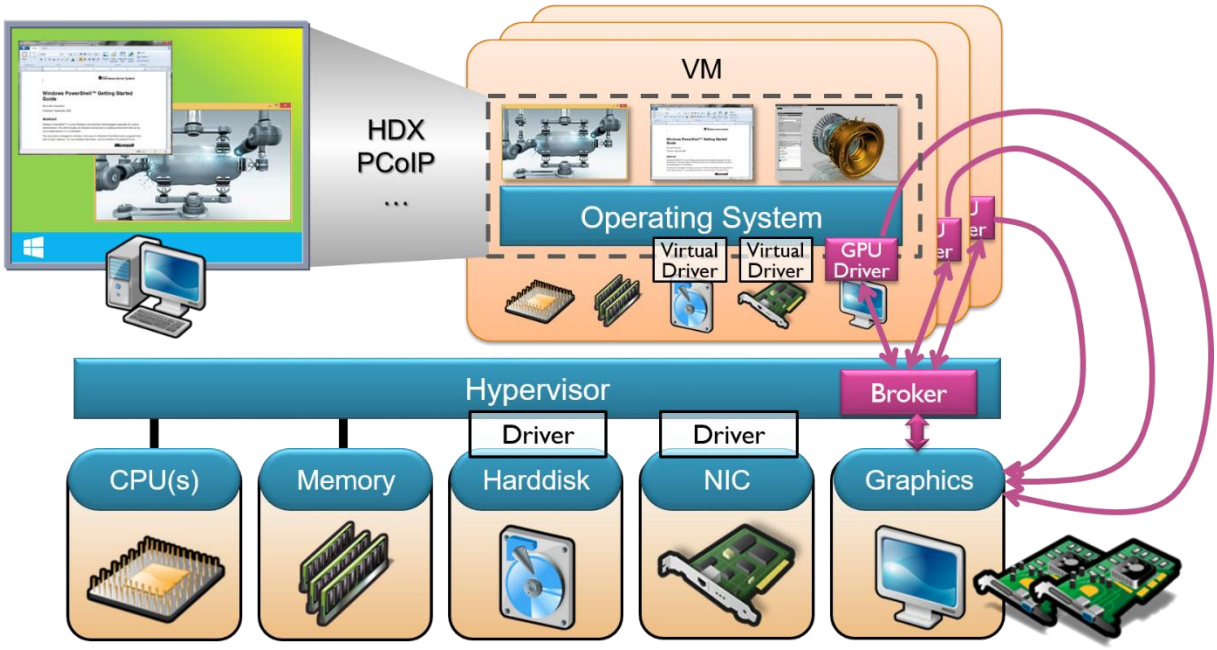
- Cost effective platform to deliver 3D graphics for virtual desktops/apps

Challenges:

- Some 3D application may not work in a multi-user environment
- The solution doesn't always meet the requirements for high-end graphics where users will view, create, manipulate, render complex 2D/3D graphics
- The GPU, particularly the GPU framebuffer, can become a performance bottleneck as many users will share the same GPU

GPU Virtualization for VDI - vGPU

GPU Virtualization for VDI or hardware virtualized GPU offers the benefit of GPU sharing and gives the virtual machine via de native graphic driver access to the GPU. Even though implemented differently by each vendor, this technology allows to share an AMD, Intel or NVIDIA GPU by multiple VMs.



In the case of NVIDIA GRID each VM is configured with a vGPU profile with a configurable static vRAM size and the VM will share GPU processing power. Citrix XenServer and VMware ESXi/vSphere are the virtualization hypervisors in the market that supports NVIDIA GRID vGPU.

Benefits:

- Full support of 3D graphics formats such as OpenGL and DirectX
- Depending on the vGPU profile the solution can deliver high-end graphics performance combined with highest-user density for knowledge workers or operators

Challenges:

- Lower overall VM density per GPU as compared to GPU Sharing for Virtual Desktops

- No OpenCL and CUDA API support today (Note: CUDA added with GRID v1.1)
- No VM migration, limitation of NVIDIA GRID architecture in current generations of GPUs

Application Vendor Support

GPU virtualization and pass-through GPU for VDI allow applications to use vendor-specific graphics drivers. ISVs such as Autodesk, Solidworks, Siemens, ESRI and PTC certify their solutions for particular graphics drivers.

- Application certification for NVIDIA GRID can be found here: http://www.nvidia.com/content/cloud-computing/pdf/GRID_Certifications.pdf.
- Application certification for AMD FirePro can be found here: http://support.amd.com/en-us/download/workstation/certified.

How to Choose the Right 3D Graphics for Virtual Desktop Solution?

What is the right 3D graphics for virtual desktop solution and how do you make the choice for that solution? First of all, when you don't know the requirements from end-user, application or IT perspective the easiest and best way to determine the right solution is flipping the coin.

It is key to classify application usage, perceived performance as well as IT and user acceptance. The diagram below is intended to help you when selecting the right model for you as it highlights the different 3D graphics concepts and its characteristics.

3D Graphics for Virtual Desktops	High Performance	3D API Support- (OpenGL, DirectX)	Cost Effective	Application Compatibility
Software 3D Graphics	★	★	★★	★
Bare Metal OS	★★★★	★★★★	★	★★★★
GPU Pass-through	★★★★	★★★★	★	★★★★
GPU Virtualization - vGPU	★★★	★★★	★★★	★★★
GPU Sharing for VDI	★★	★★	★★★★	★★
GPU Sharing for RDSH	★★	★★	★★★★	★★

Client Platforms Accessing 3D Graphics

In today's world of Consumerization of IT (CoIT), Bring Your Own and Corporate Owned Personally Enabled (COPE) scenarios there is not one single platform anymore. Today's reality is laptops, hybrid laptop/tablet devices, Ultrabooks, zero clients, thin clients, desktops and workstations running Windows, OSX, Android or ChromeOS. On the mobile side there are android, iOS, Windows Phone, Blackberry and MozillaOS as native platforms and HTML5 as a more generic web application delivery platform.

What is the best platform for accessing 3D graphics applications and/or desktops? Is it a Google Chromebook with HTML5? Is it a zero client with smartcard integration? Is it a Windows 10 tablet with touch screen? Is it a MacBook Pro accessing AutoDesk Revit? Or is it a Windows 7 corporate laptop accessing PTC Creo? The simple answer is that the best platform for accessing 3D graphics

applications and/or desktops does not exist. It all depends on perceived performance, user experience, required functionality such as USB support, multi-monitor support, client printing, network scenarios and many others. So the main question is do you know what your business consumers, your partners and co-workers expect from the solution and do you know the impact on endpoint choice and platform accessing 3D graphics applications?

Guest OS for 3D Graphics

The guest operating system is the platform being used for delivering Windows or Linux 3D graphics applications. This OS is installed inside the virtual machine, physical server or professional workstation and is responsible for running the 3D graphic application.

What is the best guest OS for delivering 3D graphics applications? The main question to answer is what are the application requirements? Does the application require a Linux or Windows Client OS, can the application leverage a 64-bit operating system? Is VDI or RDSH being used? Do you want to deliver the Windows application as a service through a service provider and what is the impact on licensing? In such as case a Windows Server OS used in a single-user or personal mode can be a good solution.

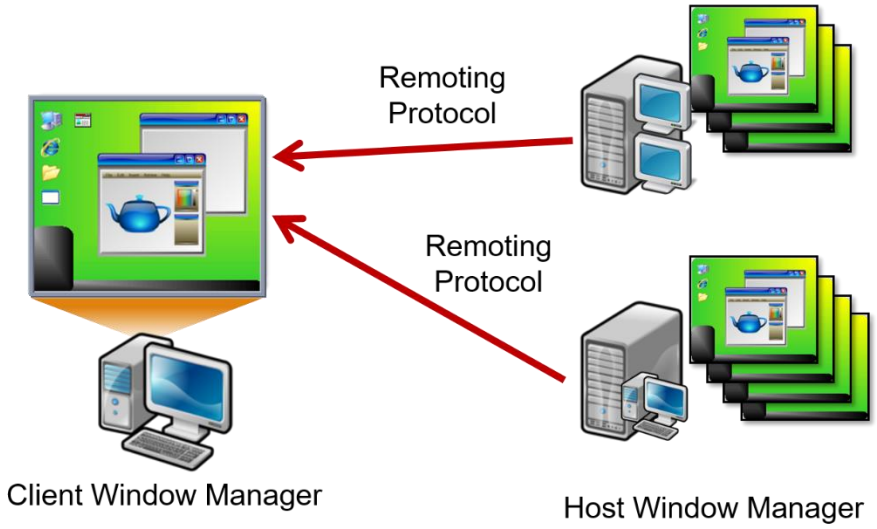
Keep in mind that the majority of 3D graphics for virtual desktop solutions are focused on delivering Windows applications and desktops to users. The VDI products offered by Citrix, Microsoft and VMware are examples. Remoting protocols and corresponding solutions such as Citrix HDX, HP RGS, Nice DCV and VMware Horizon do support both Windows and Linux guest operating systems.

Remoting Protocols Turned Inside Out

This chapter highlights the concepts behind remote access to contents and display information that are based upon different common graphics formats.

Graphics Remoting Fundamentals

User experience remoting has been around since the 60s of the last century, initially focusing on text-only remoting across serial lines. Modern graphics remoting goes far beyond such a basic model. It allows rich Windows/Linux applications and their graphical output device to be separated by a remoting boundary. It facilitates user interaction with a remote computer system by using a remoting protocol to transfer graphics display data from a host system to the user sitting in front of a client system. User input is transported from the client to the host and replayed there.



Modern versions of remoting protocols, such as Microsoft RemoteFX, Citrix ICA/HDX or Teradici PCoIP try to improve performance by taking advantage of physical Graphics Processing Units (GPUs) on the remoting host. In addition, the remoting protocols identify the client capabilities at connection time and constantly analyze the network conditions throughout the entire remoting session time in order to adjust communication settings dynamically. This all helps to make the desktop remoting environment self-adaptable and provide the best performance possible.

When looking at the impact network conditions have on the performance of remoting protocols, most people believe that bandwidth is the most significant factor. While this is true for networks with a low bandwidth profile, it is different for networks with more than 2Mbps per remote user session. In such cases, latency and packet loss become the limiting factors. User experience will typically degrade when latency is more than 50ms and it will be very challenging when latency exceeds the 200ms threshold. Packet loss should be below 1% for a good user experience. However, there are new remoting protocol variants that were specifically designed to perform well in high latency/packet loss scenarios. Citrix ICA/HDX-Framehawk is such a technology. In this context you should keep the Framehawk bandwidth requirements in mind, with Framehawk typically consuming more bandwidth than normal HDX.

NOTE: Remoting protocols typically don't limit themselves with respect to available network resources. In other words, if one remote user session requires high bandwidth for rich multimedia content, it may well consume up to 100Mbps if such bandwidth is available.

Remoting Protocol Features

Remoting protocols run on top of the Internet Protocol (IP), using Transmission Control Protocol (TCP), User Datagram Protocol (UDP) or a combination of TCP and UDP for different aspects of remoting. While older remoting protocols only used TCP, the modern ones use UDP for the graphics remoting aspect.

TCP is a connection-oriented protocol providing high reliability through error checking, congestion control and a built-in mechanism that rearranges data packets in the order specified. It also guarantees that all data remains intact in the packets transferred. But all this makes TCP relatively heavy-weight, significantly reducing graphics remoting performance on low bandwidth and high latency/packet loss networks.

UDP is a connectionless protocol, allowing a program to send individual data packets to another program. There is no error checking and no guarantee that all packets were delivered in the right order. Only if a data packet arrives at its destination it is checked for integrity. This makes UDP light-weight and much faster than TCP, but at the risk of losing chunks of data. The receiver program needs to be prepared to deal with this kind of data loss.

But there is more in a remoting protocol, in particular when it comes to extensibility. The concept of virtual channels provides a way to establish separate streams of data communication while taking advantage of the remote session communication already established. Many remoting protocols use virtual channels to add functions that allow a strict separation from the core features or are not yet specified in the protocol. They represent a platform that future developments can be based on without having to modify the communication methods between host and clients. Examples for virtual channel use cases are joint client and server clipboards or redirecting print jobs to local client printers.

Other notable remoting protocol features include bi-directional audio transmission, client side caching mechanisms, session reconnect after a temporary loss of connectivity, device redirection, multi-monitor support, location awareness and support of Unified Communications. They all are relevant for an acceptable user experience.

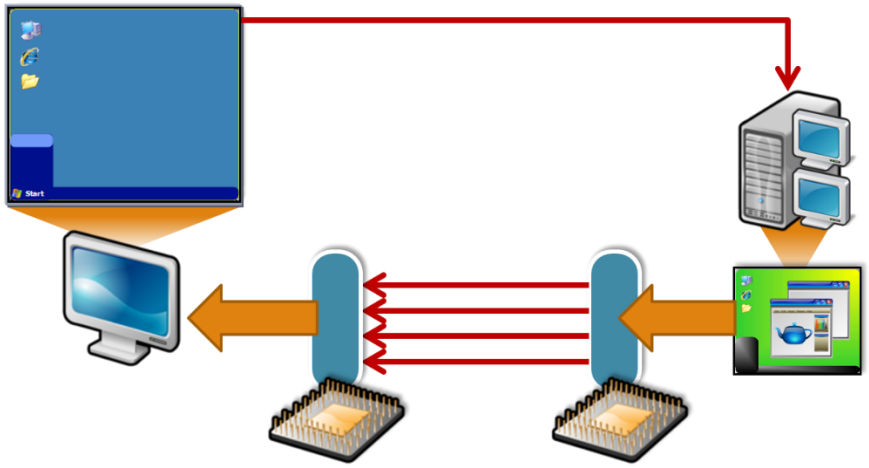
In most cases, transport security is built into the remoting protocols. Typically, they are taking advantage of the Transport Layer Security (TLS) or Datagram Transport Layer Security (DTLS) communication protocols to prevent attack scenarios such as eavesdropping. These protocols use X.509 certificates to encrypt data flowing between host and client.

Client Side Rendering versus Host Side Rendering

In a graphics remoting environment, the Windows desktop including its applications is rendered in a different way compared to traditional, local PCs. But what exactly is “rendering”? The most common definition is as follows: If an image described in an abstract model using vector graphics and bitmap primitives is converted into a raster image for output on a screen the term rendering is used. In graphics remoting environments, rendering may happen either on the client side or on the host side.

When graphics primitives are transmitted from the host to the client and rendered on the receiver side, this is referred to as client side rendering. When multimedia data streams are involved, the preferred term is multimedia redirection (MMR). The advantage of this remoting method is that it can greatly reduce host side CPU impact. In addition, redirecting media typically works well in

constrained network and due to the nature of the transmitted objects it has a lower bandwidth profile than host side rendering.

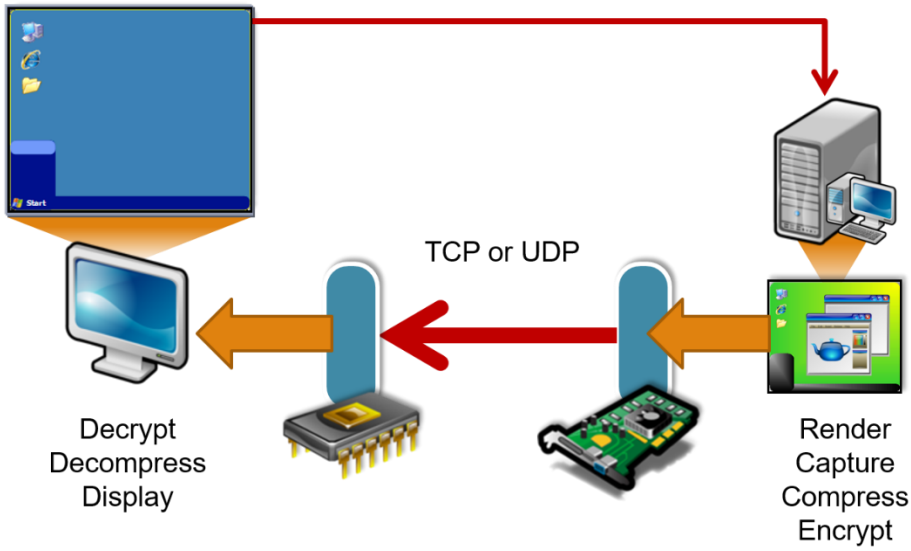


But client side rendering also has many disadvantages: The breadth of features can be very dependent of the client operating system, some forms of media may fall back to server side rendering, client rendered media may require some buffering time, and client side resource requirements may be significantly increased. In addition to that, client side rendering introduces audio/video codec dependencies and possible security vulnerabilities through outdated client side components. If the client is a mobile device, local rendering may be not desired as it consumes too much battery power.

Another significant challenge of client side rendering is the synchronized playback of audio and video in film or animation sequences. This is also referred to as “lip sync”, which is a common term for matching lip movements with spoken vocals. Dedicated codecs and processing pipelines for the separate audio and video data streams in combination with network delays may result in A/V sequences that are out of sync.

Until recently, client side rendering was regarded as the preferred Windows desktop remoting method, providing superior performance and user experience. But modern Windows desktops and applications may use various graphics formats which makes it hard to implement associated rendering algorithms across all popular client platforms. As a result, modern remoting protocols are typically not based on client side rendering anymore.

When graphics primitives are rendered on the host device instead of on the client device, this is called host side rendering. With this rendering model, all graphics types can be by transmitted as a highly compressed bitmap images to the endpoint device in an adaptive manner. Ideally, the host computer takes advantage of one or multiple physical GPUs to improve graphics performance.



The advantage of host side rendering is that it generally is client OS independent and also independent of client hardware. More advantages are that any form of mixed media content can be rendered, media content playback begins immediately, only fairly low client side resources are required and audio/video playback is provided regardless of client side codecs. In addition, modern business cases are built around aggregating technology and budget in datacenters, making host side rendering the natural choice.

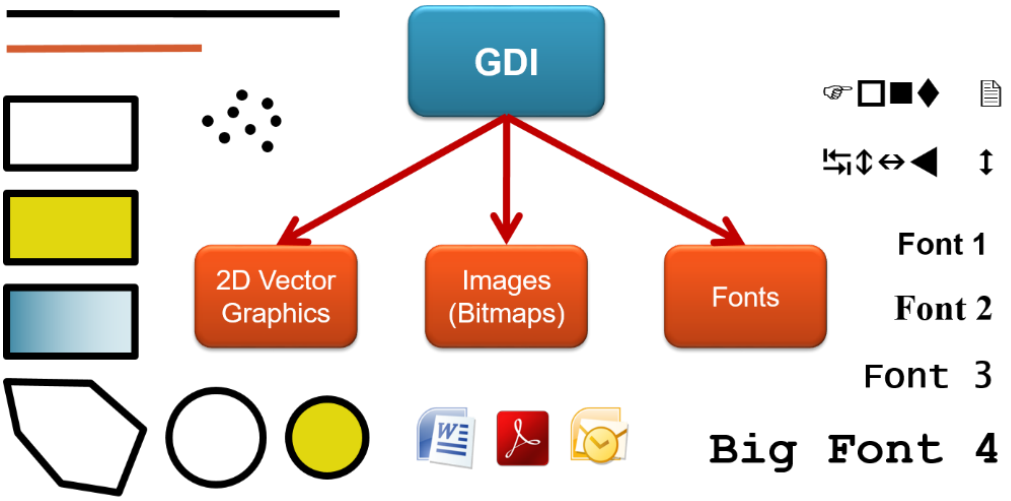
Disadvantages are that host side rendering may generate a very high host CPU impact, it can have A/V sync issues in constrained networks (depending on the protocol's implementation) and under certain circumstances it may require a higher bandwidth profile if compared to a client or hybrid approach.

As mentioned earlier, Windows desktops and applications are not only built on top of one single graphics format. As a matter of fact, a typical modern Windows system uses multiple graphics formats side by side for composing the desktop. It is important to understand nature and purpose of the individual graphics formats as some of them are treated differently in the different remoting protocols. The following paragraphs introduce the most important graphics and multimedia formats used for Windows desktop composition.

GDI Remoting

Graphics Device Interface (GDI) is an application programming interface (API) that was developed by Microsoft in the 80's and 90's of the last century. It is responsible for defining and rendering graphical objects on Microsoft Windows. GDI rendering or rasterization is the task of taking an image described through vector shapes or primitives and converting it into a raster image consisting of pixels or dots for output on a GDI-compatible video display or printer. Such a rasterized image can also be stored in a standard bitmap file format.

Typical GDI primitives are lines, rectangles, ellipses, arcs, Bezier splines, filled areas, bitmaps and fonts. A significant capability of GDI is its indirect method of accessing the underlying hardware by drawing on a Device Context (DC). A DC represents an abstraction layer that can be mapped to multiple physical target devices, such as screens and printers. A GDI applications sends its graphics output to the DC and then the Windows system sends the DC content to the target device.



It is important to note that GDI was designed in such a way that it can be hardware accelerated in a graphics card. Every modern PC graphics card supports the full GDI function set in hardware. During the Windows initialization phase, GDI hardware acceleration is registered in the system, significantly reducing CPU load generated by graphics output during system runtime.

In GDI there is no mechanism of synchronizing the DC with the graphics card frame buffer. As a result, GDI is not good at animating graphics primitives. There is no built-in method for double buffering animated output. In addition to that, GDI is a 2D graphics system, so it lacks the capability of rendering 3D primitives. The only 3D capability of GDI is the z-order of a window indicating its position in a stack of potentially overlapping windows. A window in the z-order list overlaps all other windows that are closer to the bottom of the z-order.

Most traditional Windows applications are built on top of GDI. The application windows and dialog boxes consist of basic GDI window objects, such as borders, title bars, captions, control boxes, scroll bars, menu bars and icons. These window objects are built of even more elementary GDI primitives, creating a hierarchical system of GDI primitives and window objects. A message pipeline allows an application to send graphical output to the client area of one of its windows or dialog boxes.

When Windows XP was introduced, GDI was extended by the GDI+ subsystem. GDI+ added more 2D graphics features and the support of modern graphics file formats. The Microsoft .NET Framework includes a managed GDI+ interface through the System. Drawing namespace. Substantial portions of GDI+ are not hardware accelerated.

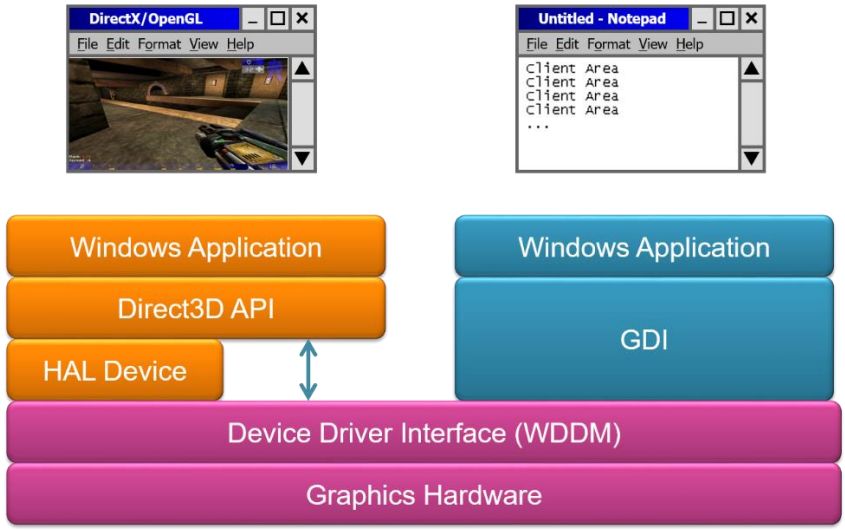
The first designs of the Microsoft Remote Desktop Protocol (RDP) adopted many concepts of GDI and GDI+. The general idea was that remoting client and host negotiated their common set of GDI capabilities during connection, allowing for an advanced level of client side rendering. If the resulting session properties did not include the support of GDI elements like bitmaps, pointers, fonts, brushes, glyphs and standard windows, the communication fell back to pixels. This was referred to as “screen scraping”, where all pixels of a rendered desktop were “scraped” from the host video memory and then transmitted to the client.

Today, both GDI and GDI+ are not as relevant as they used to be in the Windows XP times. They were replaced by more modern graphics formats such as DirectX and Windows Presentation Foundation. As GDI falls in popularity, so does the relevance of client side rendering.

DirectX Remoting

DirectX is a collection of application programming interfaces (APIs), covering different aspects of multimedia formats. These include Direct3D (D3D) for 3D graphics, DirectDraw (D2D) for 2D graphics and audio/video APIs called DirectMusic, DirectPlay and DirectSound. DirectX was introduced to compensate for the shortcomings of the Graphics Device Interface (GDI), in particular in Windows games. It is directly connected to display drivers and gets better results at rendering than GDI.

When Windows applications are using Direct3D they write to a 3D surface. In Windows Vista and later versions, Direct3D uses the Windows Display Driver Model (WDDM) to share the 3D surface with the Desktop Window Manager (DWM). DWM then uses this surface directly and maps it on to the window desktop. One of the central element of the WDDM graphics architecture is the new video memory manager. It supports the virtualization of graphics hardware for a range of applications and services like the Desktop Window Manager.



If an adequate graphics card is available, Direct3D can use it for hardware acceleration of the entire 3D rendering pipeline or part of it. The advanced graphics capabilities of the underlying 3D graphics hardware are exposed through Direct3D. Combining such capabilities with the other DirectX technologies allows for advanced multimedia scenarios.

Over the years, Direct3D became a de-facto standard for many software vendors when developing software applications for visualization, games and other high-end graphics tasks. Direct3D 9, Direct3D 10 and Direct3D 11 are only available on Windows Vista and later due to the fact that they require the Windows Display Driver Model.

NOTE: Microsoft released Direct3D in 1995, which became a competitor of OpenGL. In December 1997, Microsoft and Silicon Graphics initiated a joint effort with the goal of unifying the OpenGL and Direct3D interfaces.

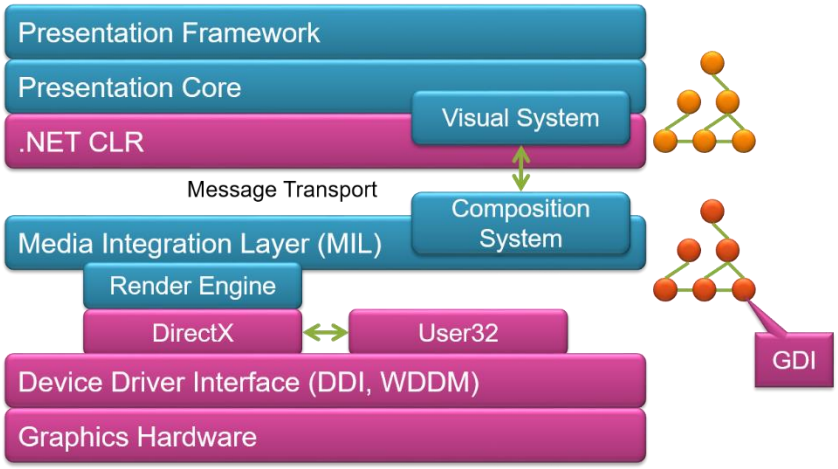
Remoting DirectX is typically done on the host side, resulting in very poor performance when using the older remoting protocols. However, there are forms of client-side DX remoting that are enabled by both Microsoft (Aero Glass Remoting in RDP) as well as Citrix XenDesktop (Desktop Composition Redirection). These forms of DX remoting even work well in decent WAN conditions, but they are designed to support basic DX remoting and are not designed to support high-end DirectX graphics remoting. By redirecting DirectX calls to the client device, you can reduce the server load on the

hosted desktop platform thereby scaling more users. The downside of this approach is a more limited set of client hardware / OS support (though fallback to server side rendering occurs in those cases).

WPF Remoting

Windows Presentation Foundation (WPF) was introduced with Windows Vista as part of the .NET Framework 3.0. It was designed to be the successor of GDI for the standard graphical user interface of Windows desktops and applications.

WPF is completely built on top of DirectX, enabling modern UI features like transparency, gradients and transforms. One of the major WPF design goals was to provide a consistent programming model for building applications. In particular, it separates the user interface from business logic by providing a markup language (XAML) to define UI elements and relationships with other UI elements.



WPF applications allow for a variety of runtime options. They can either be deployed as stand-alone desktop programs or hosted as an embedded object in a website. Independently of the deployment method, WPF supports a basic set of common user interface elements for 2D and 3D rendering, including vector graphics, runtime animation, and pre-rendered media. These elements can then be linked and manipulated based on various events, user interactions, and data bindings. One such element is the 2D surface that is generated by all open GDI applications. This GDI surface is mapped on the WPF-based Desktop Window Manager. As a result, this concept for desktop composition allows to combine 2D and 3D elements seamlessly, including moving bitmaps, transparency and anti-aliasing. One of the biggest advantages for GDI applications when used in the WPF context is that the application logic doesn't need to re-render graphical content when the application window comes to the foreground after it was covered by another window before.

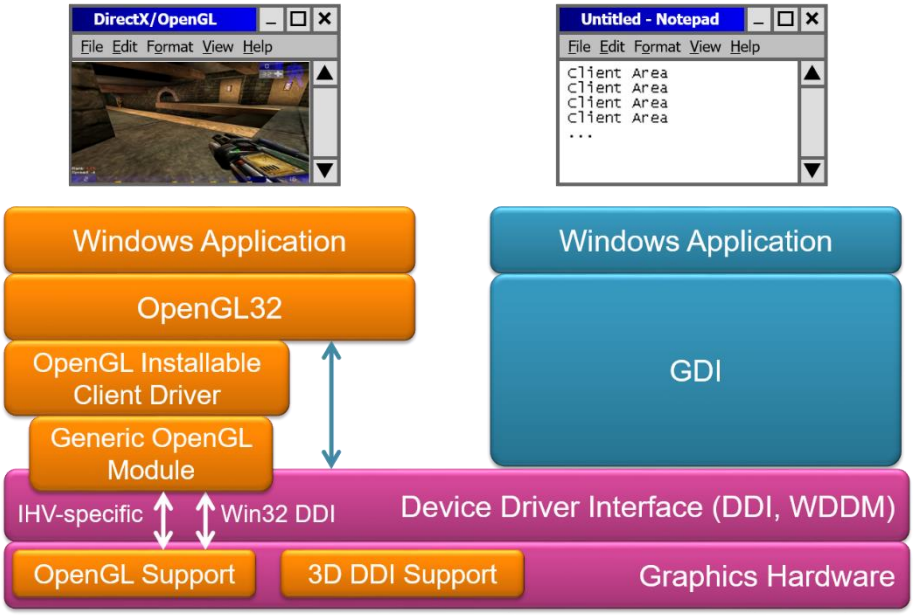
Remoting WPF is identical to remoting DirectX, which means that it is typically rendered on the host side and can be GPU accelerated.

OpenGL/WebGL Remoting

Open Graphics Library (OpenGL) is a multi-platform application programming interface (API) widely used in 2D and 3D applications. It was developed by Silicon Graphics Inc. in the 90s and is managed by a non-profit technology consortium, the Khronos Group, since 2006. The OpenGL API consists of over 250 different function calls to draw complex three-dimensional scenes from simple primitives. It is capable of interacting with a physical GPU for hardware-accelerated graphics rendering. Version 1.0 of OpenGL was released in January 1992. New versions incorporate a number of extensions,

defining a new set of features which all conforming graphics cards must support. OpenGL 1.1 was released in 1997 and is the supported version in Remote Desktop Protocol (RDP) 8.1 including RemoteFX. OpenGL 1.2, 1.3, 1.4 and 1.5 followed in 1998, 2001, 2002 and 2003 respectively. OpenGL 2.0 introduced a Shading Language in 2004 and was succeeded by version 2.1 in 2006. OpenGL 2.1 is the supported version in VMware vSGA.

OpenGL 3.0 was released in 2008 and it introduced a deprecation mechanism for API revisions. OpenGL 3.1, 3.2 and 3.3 followed in the years until 2010. HDX 3D Pro is fully compatible to OpenGL 3.3 and earlier. The design goal of OpenGL 4.0 was to support DirectX 11 compatible hardware. It was also released in 2010 alongside version 3.3. New NVIDIA and AMD graphics cards led to versions 4.1, 4.2, 4.3 and finally 4.4. In general, OpenGL releases are backwards compatible. Most graphics cards released after the release date of a particular OpenGL version support those version features, and all earlier features.



In remoting scenarios, OpenGL is typically rendered on the host side, and it can greatly benefit from GPU acceleration.

WebGL is a cross-platform web standard for a low-level 3D graphics API. It is based on OpenGL ES 2.0 and exposed in Apple Safari, Google Chrome, Mozilla Firefox and Opera browsers through the HTML5 Canvas element as Document Object Model interfaces.

OpenCL Remoting

Open Computing Language (OpenCL) is an open, cross-platform standard for writing gaming, entertainment, scientific and medical software. It was designed for parallel programming of CPUs, GPUs, digital signal processors and field-programmable gate arrays while it interoperates with OpenGL and other graphics APIs. Like OpenGL, OpenCL is maintained by the non-profit technology consortium Khronos Group. It has been adopted by vendors such as NVIDIA, AMD, Intel, Apple, Qualcomm, Samsung and ARM Holdings.

Version 1.0 of OpenCL was released in December 2008, allowing applications to tap into multi-core CPUs and GPUs. OpenCL 1.1 and 1.2 followed in June 2010 and November 2011 respectively, adding functionality for enhanced parallel programming and improved OpenGL and DirectX interoperability.

OpenCL 2.1 is the latest significant version. It was released in March 2015 and is designed to further simplify cross-platform parallel programming.

Remoting OpenCL is pretty much identical to remoting OpenGL. Rendering typically happens on the sender side.

Flash Remoting

Adobe Flash is a multimedia platform to create Web advertisements, interactive games and Rich Internet Applications (RIAs), allowing the combination of still images, animation, vector graphics, audio and video. Viewing and interacting with the interactive multimedia content of a Flash application requires the Adobe Flash Player which supports the Flash Video Format (FLV). From a playback perspective, the Flash Player has a lot in common with a custom-developed video player using a proprietary multimedia format. But in contrast to a video player, Flash allows user interaction through mouse, keyboard, microphone and camera. This includes bidirectional streaming of audio and video data. A clear advantage of Flash is that the Flash Player is available free of charge as a plug-in for all major Web browsers and for many operating systems, smartphones and tablet devices.

Creating interactive Flash applications is based on an object-oriented programming language called ActionScript. Professional development tools are required to build such application. Since a couple of years, the introduction of new or updated Flash applications is declining on websites. However, there are still many Flash applications in use.

Remoting Flash is typically done on the host side, resulting in very poor performance when using the older remoting protocols. Only Citrix and Dell implemented a technology called Flash Redirection which relies on an existing Flash Player on the client side. Flash Redirection improves usability and reduces bandwidth requirements significantly, but also requires a client that is capable of installing and maintaining a Flash Player. There are some negative aspects of performing client side Flash remoting such as a dependency of keeping the server and client Flash versions up to date as well as some security related risks when redirecting Flash to the client system (i.e. a server hosted Flash control could infect a client PC since the Flash content is rendered on the client).

Silverlight Remoting

Microsoft released Silverlight in 2007 with the ambition to replace Adobe Flash as the most popular application framework for interactive multimedia applications. Now in its fifth version, the Silverlight runtime environment is available as a plug-in for Web browsers running under Windows and Mac OS X. In addition, Silverlight is one of the Windows Phone development platforms, even though the resulting applications can only run natively on Windows Phone but not from Web pages in Internet Explorer on Windows Phone, Windows Mobile or Windows RT.

Like with Adobe Flash, Silverlight remoting is done on the host side. There is no technology available today that allows redirecting Silverlight content even if a Silverlight runtime is installed on the client side. As a result, Silverlight performance is degraded when using the older remoting protocols. Note: Adobe dropped support for flash with Linux.

Audio/Video Remoting

Playing back audio and video streams is very easy on Windows systems. The built-in Media Player supports a number of multimedia file formats, such as WAV for audio and WMV or AVI for video. RealVideo and Apple QuickTime can be seen as competitors.

Most video files consist of two or more types of files objects. One is the media container which describes the overall structure while another represents the compression/decompression (codec) algorithm used inside the container.

Here is a short overview of the most popular file formats and codecs:

- MPEG-1 and MPEG-2 files are based on a compression format standardized by the Moving Picture Experts Group
- MP4 is a video format that enhances the MPEG standard by the support of video/audio "objects", 3D content, low bitrate encoding and Digital Rights Management. In addition, it allows the separation of audio and video tracks. Video is compressed with MPEG-4 and audio uses AAC compression
- Audio/Video Interleave (AVI) was developed by Microsoft and most commonly contains MPEG or DivX codecs even though it can contain almost any codec
- The Windows Media Video format (WMV) was also developed by Microsoft. Initially designed for Internet streaming, it is now a common format used for video playback
- QuickTime was developed by Apple. The format contains one or multiple tracks containing video, audio and other objects
- The RealMedia format was created by RealNetworks, allowing media streaming of both audio and video data
- The RealVideo format was developed by Real Media with a low bandwidth use case in mind

The most popular video codec today is H.264/MPEG-4 Part 10 or AVC (Advanced Video Coding), released in May 2003. It is a block-oriented motion-compensation-based video compression standard, also used as one of the video encoding standards for Blu-ray discs. H.264 can be implemented in hardware which allows modern GPUs to accelerate video compression and decompression on hardware level. By default, H.264 compression is lossy, but it can be configured to be visually lossless, which is an important aspect for use cases such as medical imaging.

Many remoting protocols try to redirect audio and video data streams to the client if on session connection the required codec packages can be found there. If a video stream is decoded in host video memory and re-encoded for remoting the content, this comes with substantial resource requirements.

NOTE: Remoting protocols implementing host side rendering typically use H.264 to encode and compress the Windows desktop into video frames before sending it down the wire. Taking advantage of the H.264 hardware encoder on the GPU improves remoting performance significantly.

HTML5 for Remoting

Hypertext Markup Language version 5 (HTML5 for short) is the advanced version of HTML, used for presenting text and multimedia content in the World Wide Web. An important new feature in HTML5 is the way it embeds graphics, audio, video and interactive documents, allowing to create Web applications. New syntactic features provide a simpler way to manipulate multimedia and graphical

content without the need to install additional browser plugins. The latest versions of all modern browsers, such as Internet Explorer, Firefox, Chrome, Safari and Opera, support HTML5 to a certain extend. A public HTML5 test site (http://html5test.com/) creates a HTML5 compatibility score.

An interesting aspect of HTML5 is its capability to deliver remote desktops via native browser components. This requires WebSockets to create a TCP connection channel for continuous transmission of data throughout the entire remote session lifespan as opposed to the connectionless communication via HTTP. In addition, a native HTML5 element called Canvas is needed that has the ability to control single pixels discretely and output complete raster images at high speed. In combination, WebSockets and Canvas provide the necessary mechanisms to connect to a remoting host from a browser and render the 2D graphics output dynamically. As a result, the browser receives the remoting data through WebSockets and uses the Canvas element to draw the desktop. A gateway component or driver between the host and the browser is required to render the content received through the remoting protocol and convert it into a binary data stream of desktop frames retransmitted through WebSockets.

Among industry experts and market analysts, remoting clients based on HTML5 are regarded as a top candidate for the preferred future technology. But at this stage it is too early for declaring HTML5 as the successor of native remoting clients.

CUDA in Graphics Remoting Environments

The Compute Unified Device Architecture (CUDA) is NVIDIA's parallel computing platform and programming model. CUDA was designed to be used with NVIDIA GPUs, giving developers direct access to instruction set and memory of the graphics chip. Use cases are not only limited to high-end graphics. CUDA GPUs can also be used for general purpose processing, an approach called GPGPU (General Purpose Computation on Graphics Processing). The advantage a GPU has over a CPU is that GPUs are optimized for running a big number of concurrent processes or threads side by side, resulting in superior overall performance. This makes GPGPU extremely powerful for scientific calculations and cryptography.

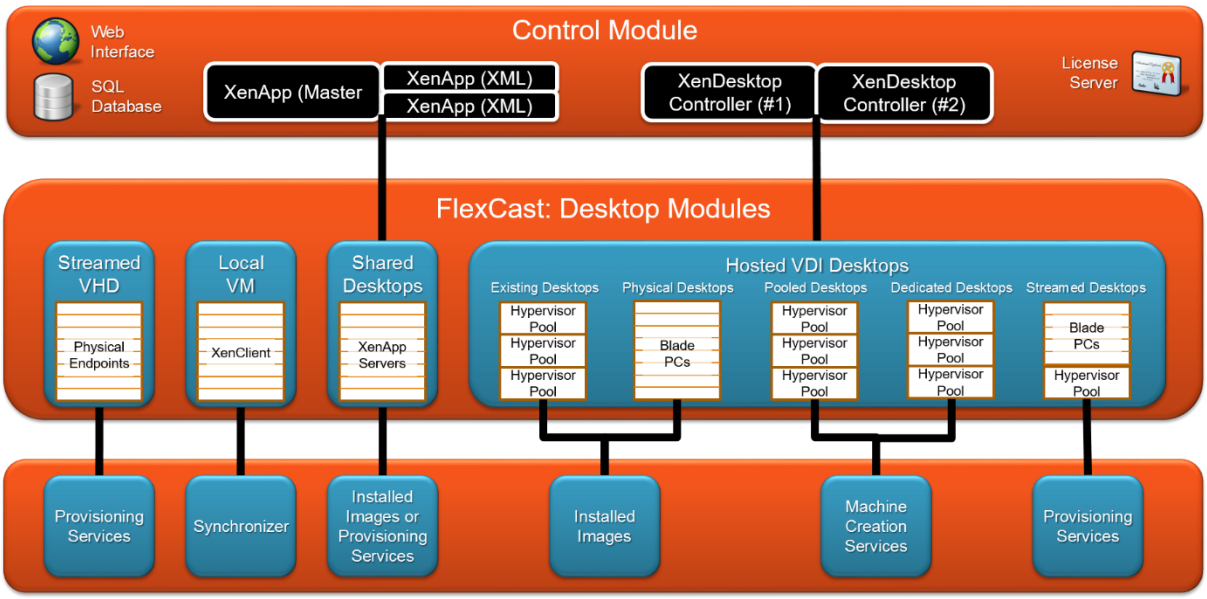
NVIDIA offers CUDA-accelerated libraries and extensions to standard programming languages such as C and C++, allowing developers to take advantage of the GPU's capabilities. Vendors like Adobe or Autodesk use such features to accelerate their 3D CAD/CAM programs. CUDA-enabled applications run if there is a physical NVIDIA GPU present, or when GPU pass-through is used. Remoting such applications also benefits of the hardware acceleration on the GPU.

VDI Vendors and Their Solutions

This chapter will give you an overview of the major vendors and their solutions to deliver 3D graphics for virtual desktop.

Citrix XenApp and XenDesktop

Citrix has a long history in Server-Based Computing (SBC) with its XenApp product. To expand the market, Citrix focused on using a broad set of virtualization technologies to create XenDesktop, an end-to-end solution for providing virtual desktops and apps for a broad set of use cases. XenDesktop includes the ability to host personal, pooled, and dedicated virtual machines (traditionally known as VDI,) but also enables cost-effective and scalable hosted-shared environments with powerful personalization tools, locally executed desktops with XenClient (a Type-1 hypervisor) and streamed VHD (real-time streamed OS at boot). Finally, XenDesktop also allows GPU-enabled PC blades and GPU assisted Hypervisors for providing rich, pixel-perfect 3D professional graphics support. All these different options are all part of the Citrix Flexcast Management Architecture (FMA).



However, not every user needs a full virtual desktop. Increasingly, the demands for extending Windows apps to tablet and smartphone form factors or the need to provision virtual apps on-demand to BYO laptops have increased the need for an integrated application virtualization and delivery infrastructure. XenDesktop includes these capabilities to support enterprise workers who may roam between device types and networks but need a consistent work environment that is easy to use, centrally managed and secure.

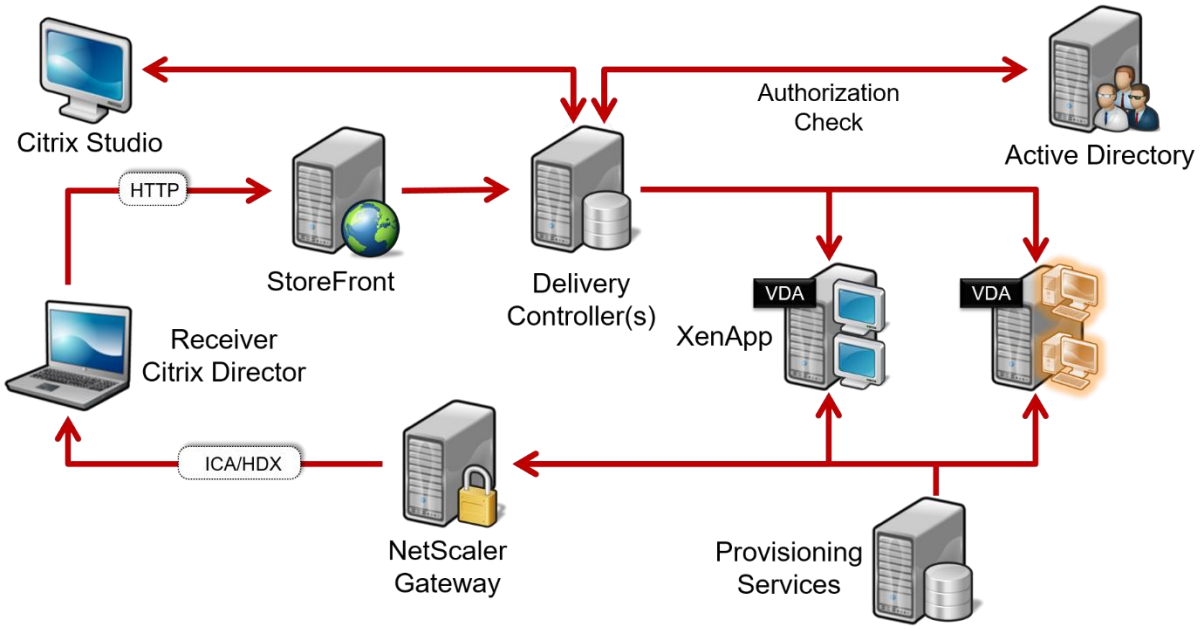
Architecture

The architecture of XenDesktop, today, consists of multiple components. This makes the initial setup of XenDesktop more complex than some other solutions, but is designed for flexibility and scale. The tasks of the connection broker or the "Delivery Controller" are:

- Authenticates users
- Manages assembling the user's desktop environment
- Brokers connections between the user and his virtual desktop

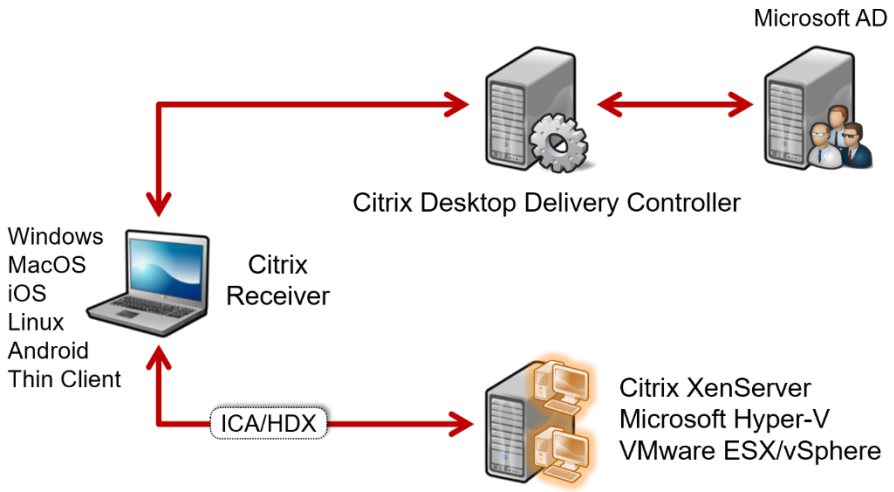
Other components of the XenDesktop infrastructure are: Virtual Infrastructure: the hypervisor (XenServer, Microsoft Hyper-V or VMware vSphere/ESXi), licensing server, database server, Provisioning Server, Storefront, Director and NetScaler Gateway for secure remote access.

Introduced in XenDesktop 7 is a single architecture to host virtual machines, remote PC and hosted shared desktops. XenApp 7.5 delivers shared desktops through a Remote Desktop Services based Windows Server OS. The XenApp and XenDesktop infrastructure is hereby capable to include access to remote or streamed applications on a Windows Server OS.



A XenDesktop agent is required in the guest VM or Blade PC. The agent enables direct connections between the endpoint and the user's virtual desktop. On the endpoint the Citrix Receiver or Receiver for HTML5 can be used to provide simple access from any device, anywhere, to virtual desktops and applications. The Citrix Receiver provides the most comprehensive experience, while the Receiver for HTML5 has less endpoint requirements as well as capabilities.

The protocol used for XenDesktop (and XenApp) is the ICA protocol which has some special features. The ICA protocol performs better on limited bandwidth and higher latency than the RDP/RemoteFX protocol. Citrix HDX technologies were introduced in XenDesktop 3. HDX includes several elements that improve user experience, such as VoIP and webcam support USB peripheral support, 3D support, enhanced audio and optimization for use over WAN.



Licensing

Each license type includes, next to the XenDesktop Controller (the broker) the right to use certain components and features of Citrix XenDesktop: VDI desktops, Hosted Shared Desktops, XenServer, Provisioning Server, NetScaler Gateway, EdgeSight, etc. VDI edition is the most basic version with a partial but still strong feature set for pure VDI deployments. Platinum is the most advanced edition and delivers the complete feature set.

NOTE: With respect to 3D graphics support, you must have the Enterprise or Platinum Editions of XenApp/XenDesktop to leverage the HDX3DPro feature set.

	XenDesktop Editions			XenApp Editions		
Features	VDI	Enterprise	Platinum	Advanced	Enterprise	Platinum
App Hosting	X	√	√	√	√	√
VM-Hosted Apps	X	√	√	X	√	√
App-V for Offline-Apps	X	√	√	X	√	√
Hosted Shared Desktop	X	√	√	√	√	√
XenClient Enterprise	X	√	√	X	X	X
Pooled VDI Desktop	√	√	√	X	X	X
Dedicated VDI Desktop	√	√	√	X	X	X
Physical Desktop	X	√	√	X	X	X
Remote PC	X	√	√	X	X	X
HDX Seamless Local Apps	X	X	√	X	X	√
Personal vDisk	√	√	√	X	X	X

An overview of the different XenApp and XenDesktop versions and its features can be found here: https://www.citrix.com/go/products/xendesktop/feature-matrix.

Citrix and Hardware Accelerated Graphics for Desktop Virtualization

Citrix's XenDesktop and XenApp support for 3D graphics for Virtual desktops timeline:

- **2009**: Availability of XenDesktop HDX 3D Pro with Deep Compression
- **2011**: XenServer 6.0 supports GPU pass-through
- **2012**: HDX3DPro increased fps via NVIDIA GRID. Improved compression and bandwidth reduction
- **2013**:
  - XenDesktop 7.0+ GPU sharing via GRID K1/K2 interfaces
  - XenApp 6.5+ GPU sharing (OpenGL/WebGL)
  - NVIDIA vGPU released and integrated in XenServer and XenDesktop
- **2014**:
  - vGPU for XenApp
  - PVS (in addition to MCS) for vGPU provisioning
  - Generic USB Redirection on XenApp, including USB 3.0 device compatibility
    3D mouse support on XenApp (previously just on XenDesktop)
- **2015**: Linux guest OS support for XenDesktop, limited GPU support

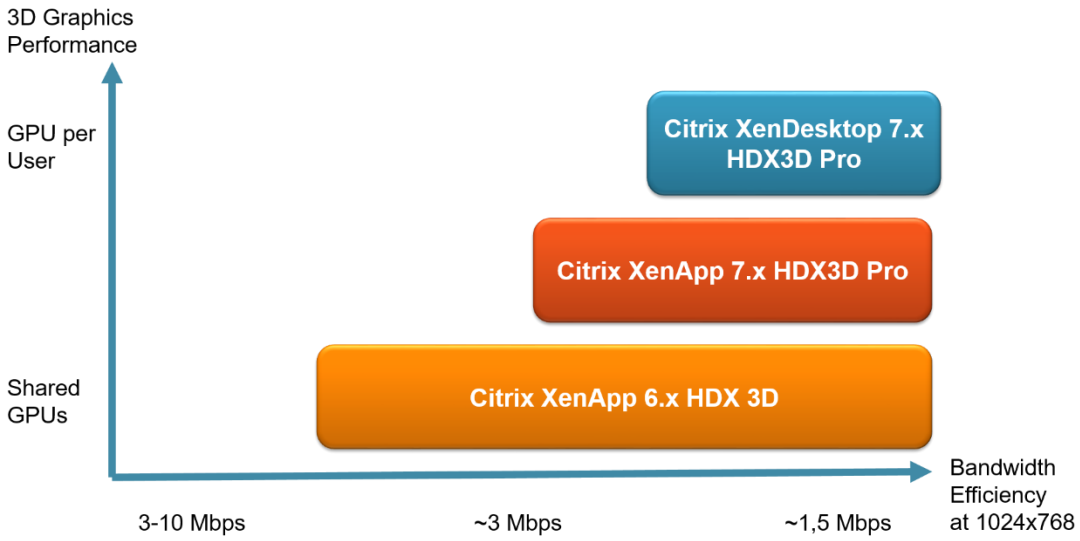


Image source: Citrix

Citrix XenDesktop or XenApp

What are the benefits of using 3D graphics with Citrix XenDesktop and Citrix XenApp?

Citrix XenDesktop is able to deliver a high-performance 3D graphics solution for the most demanding users such as design engineers who view, create, manipulate and render complex 2D/3D graphics. XenDesktop provides GPU acceleration for Direct3D, OpenGL, CUDA and OpenCL applications with H.264 based deep compression. This assures that the least amount of bandwidth is consumed. High-performance 3D graphics can be achieved because every Virtual Machine is connected to a GPU. 3D SpaceMouse support is available. Microsoft licensing for VDI, Virtual Desktop Access, can be a challenge. Especially in a Bring Your Own Device or work from anywhere scenario. Both Citrix XenServer 6.2+ and VMware vSphere 5.1+ can support Citrix XenDesktop with GPU pass-through. The

GPU virtualization technology of NVIDIA's with vGPU is currently supported with Citrix XenServer 6.2+ and VMware vSphere 6.0+ enterprise plus edition.

Citrix XenApp is able to deliver 3D graphics solutions to operators and contractors who view and manipulate 2D/3D graphics. XenApp provides GPU support acceleration for Direct3D, OpenGL, CUDA and OpenCL. Note: OpenCL and CUDA are only experimental. XenApp is installed on a multi-user Windows Server OS; applications needs to be compatible with Citrix XenApp/RDSH. A Microsoft Virtual Desktop Access (VDA) License isn't needed, instead a RDS Client Access License (CAL) is needed. Citrix XenApp with 3D graphics support is cost effective solutions because of GPU sharing. The downside is that 3D graphics performance isn't as high as Citrix XenDesktop with GPU Pass-Through or the high-performance vGPU profile.

Citrix XenServer

Citrix XenServer is an open source virtualization platform for managing cloud, server and desktop virtual infrastructures. XenServer is designed to optimize private datacenters and cloud. Citrix XenServer support both GPU pass-through and GPU virtualization via NVIDIA GRID vGPU. On XenServer, the NVIDIA driver, which represents the vGPU Manager, is installed in the control domain (Dom0) and the NVIDIA vGPU Windows driver is installed in the guest OS of the virtual machine. Although it is technically possible to pass-through multiple GPU's per VM, keep in mind that XenServer only supports one GPU per VM keep in mind that XenServer only support pass-through of one GPU per VM. Each physical GPU can be split into 1, 2, 4 or even 8 vGPUs when using NVIDIA GRID 1.0 with K1/K2.

Vendor support for different 3D graphics solutions	Software 3D-graphics	Bare Metal Client OS	Pass-Trough GPU	GPU Virtualization GRID vGPs	Shared GPU for VDI (API Intercept)	Sharing GPU for XenApp/RDSH
Citrix XenApp – Server OS (RDSH)	√	N/A	√	√	N/A	√
Citrix XenDesktop – Client OS (VDI)	√	√	√	√	X	N/A

Fra.me

Frame (formerly Mainframe2) is a secure, scalable and high-performance cloud platform that enables businesses to deliver amazing workflows and experiences to users anywhere. End-users can sign in to a Frame system from any HTML5 capable browser and have instant access to their Windows or Linux apps. No plugins, such as Java or Flash, are required.

Founded in summer 2012 as Mainframe2 and re-branded in 2015, Fra.me is based in Menlo Park, California. The Fra.me founding team is led by Dr. Nikola Bozinovic, a graphics industry veteran, and it has deep expertise in video streaming via an earlier organization, MotionDSP. Frame protocol is using H.264 as its transport layer, with significant QOS optimizations on top. Applications are streamed to any client device as a video stream. The Frame platform includes a variety of other capabilities ranging from cloud infrastructure orchestration and automated scaling to multi-region deployments on 5 continents to mobile device support. Recent advances in GPU virtualization, faster

networks and elastic clouds are allowing any user connecting to Frame to take advantage of powerful server-side resources, similar to the early days of Mainframe computing.

The vision of Frame is to make even the most demanding application available on demand to everyone with an Internet connection. To achieve this, Frame works closely with leading software vendors, including Adobe, SOLIDWORKS, and Siemens. Frame's end-user product is available as Frame for Business.

Solution Details

To ensure maximum performance and app compatibility, each active session on Frame is connected to its own Virtual Machine. Frame focuses on app delivery and not desktop delivery. Currently supported operating systems include Windows Server 2008, 2012, 2012 R2 and Linux (in beta).

Frame operates the cloud-based backplane used for brokering and resource orchestration. Customer can choose to run their workloads on any instance type on Amazon AWS or Microsoft Azure. All workloads requiring graphics acceleration currently run Amazon's g2.2xlarge EC2 instance. Support for more public clouds and for on-premises workloads is coming in 2016.

With the apps running on a VM in the cloud, rendered frames are converted to H.264 video and the application is seamlessly presented to any endpoint running a HTML5 browser or Frame terminal. For maximal performance, Frame developed and optimized its own JavaScript-based decoder, providing best-in-class end user experience supporting app delivery at up to 4K resolutions and up to 60 fps. To create a secure channel between the server and the client, Frame uses WebSockets, and there's no WebRTC involved. In addition to web terminal (HTML5), native terminals for Windows, OSX, iOS and Android are also available.

The remoting protocol is optimized to consume as little bandwidth as possible, dropping to a few kbps and 1fps when the content is still and then instantly jumping to 10Mbps+ and 60 fps when needed. This makes the solution suitable to deliver applications in LAN, WAN, 3G, and 4G scenarios.

When a user's workflow demands multiple applications, those are configured to run from the same virtual machine. This enables efficient resource utilization and features like shared clipboard. Administrators can easily set up different users with access to different apps or groups of apps. Businesses and software developers don't have to rebuild their native Windows applications for the web in order to run them on Frame. The applications, MSI, exe or ISO, are installed via a quick installation process into the platform, eliminating the need to upload massive gold master images. All administration is done for a browser-based, single pane of glass.

Frame is extensible and flexible. It comes with a rich set of integrations with cloud storage (e.g., Dropbox, Box, Google Drive, or your own), or identity services (e.g. Okta, AD/ADFS, Google). It also includes a number of advanced features such as metering, billing, and user management as well as mature set of APIs for custom development and integration. End customers, like businesses and educational institutions, pay a single bill to Frame. This covers the cost of Frame platform and all other infrastructure related costs (e.g., Amazon EC2). For more information, please visit www.fra.me

Microsoft Remote Desktop Services

With the technology originally called Terminal Services, a user is connected with a session on a Windows Server which allows for providing a full desktop session or individual programs (called RemoteApp) on the user's client device. With the launch of Windows Server 2008 R2 in October 2009, the name "Terminal Services" was replaced by "Remote Desktop Session Host" (RDSH). Windows Server 2008 R2 also introduced a new role, namely "Remote Desktop Virtualization Host". By adding this role to Microsoft's hardware virtualization platform, called Hyper-V, Remote Desktop Virtualization Host now also allows for providing users with a dedicated virtual desktop running a Windows Client operating system. With SP1 for Windows Server 2008 R2, Microsoft added RemoteFX to Remote Desktop Services (RDS). The RemoteFX technology allows for a better user experience when using Remote Desktop Services. RemoteFX allows USB redirection, high-definition video and using 3D applications over Microsoft's Remote Desktop Protocol. RemoteFX works for virtual desktops as well as for session based desktops.

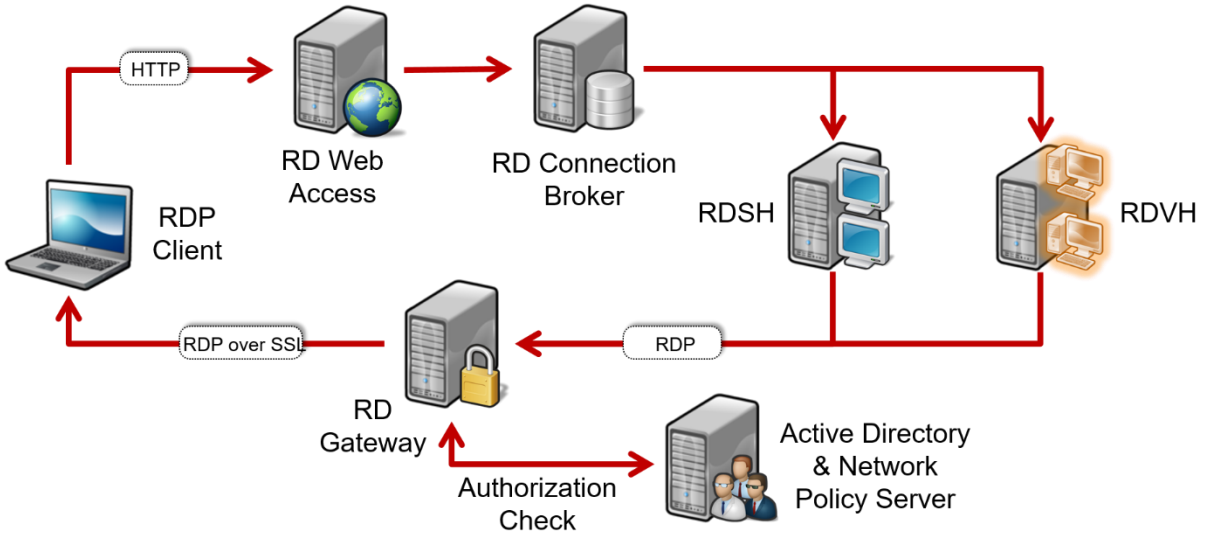
With the Windows Server 2012 release, Microsoft redesigned the complete Remote Desktop Services stack to improve the ease of deployment and management. RemoteFX features are also improved and broadened to support a new generation of Windows (8) devices. RemoteFX is optimized for WAN deployments and touch-based devices. Microsoft also delivers a built-in profile management solution for virtual desktops and session based desktops.

Architecture

Microsoft utilizes its own Remote Desktop Protocol (RDP) with RemoteFX (RFX) enhancements to provide the remote desktop or remote application to the user. For the best experience and functionality, a client device should run Windows 7 SP1 or Windows 8 with RDP8. The Microsoft Virtual Desktop Infrastructure consists of the following Windows Server 2012 roles:

- Remote Desktop Gateway (RDG). This is an optional role to provide secure access to the Microsoft Virtual Desktop Infrastructure from internet-based clients
- Remote Desktop Web Access (RDWA). This role provides access to the desktops and/or remote applications available for a specific user. After the user browses to the Web Access URL and authenticates, Web Access provides a webpage displaying the shortcuts to the resources available to this user. If the client device is running Windows 7/8 and is on the corporate LAN, the shortcuts can be also integrated in the user's Start Menu
- Remote Desktop Connection Broker (RDCB). The Connection Broker tells Web Access which resources are available to the user. The RDCB role is the broker which connects the client to the correct resource selected by the user in Web Access. The Connection Broker also contains the Remote Desktop Management Service. The Remote Desktop Management Service maintains a database with the static configuration of the deployed RDG, RDWA, RDCB, RDSH and RDVH roles, and dynamic session information of the managed RDSH and RDVH servers
- Remote Desktop Session Host (RDSH). Formerly known as a Terminal Server, RDSH provides server hosted desktops or remote applications to the client. The RDSH role is not required for a Windows Server 2012 virtual desktop infrastructure, but could be added to provide a hybrid solution
- Remote Desktop Virtualization Host (RDVH). A Virtualization Host is a Microsoft Hyper-V host with the Virtualization Host agent service installed. RDVH provides virtual desktops or

remote applications to the client. The Virtualization Host agent service manages the starting of the virtual machines or remote applications (in a virtual machine) when a user wants to connect



3D Graphics for Virtual Desktops with RemoteFX

Microsoft Remote RemoteFX delivers shared GPU access to virtual machines with Microsoft Hyper-V as a hypervisor. A DirectX 11 driver needs to be installed in the parent partition of Windows Server 2012 R2 Hyper-V to leverage shared GPU access within the virtual machines. The shared GPU functionality is also known as virtualized GPU, not to be confused by NVIDIA vGPU, with API intercept technology. Microsoft's definition of vGPU is vastly different from NVIDIA vGPU. GRID vGPU is actually a trademark of NVIDIA but Microsoft were using the term "vGPU" before NVIDIA registered it. RemoteFX in Windows Server 2012 R2 is primarily designed for fluid user experience with Line of Business applications and Windows graphical functions such as Aero. The shared GPU provides optimized graphics for DirectX enabled applications in the virtual machine. The solution is not targeted at applications which require OpenGL 1.1+, OpenCL, CUDA and users who need high-end 3D graphic performance.

The NVIDIA/AMD GPU has a dedicated amount of video RAM. Each virtual machine consumes a specific amount of video RAM. The amount is based on the maximum number of monitors and the screen resolution configured for each virtual machine. The combination of monitors and resolution determines the maximum number of virtual machines per physical NVIDIA/AMD GPU.

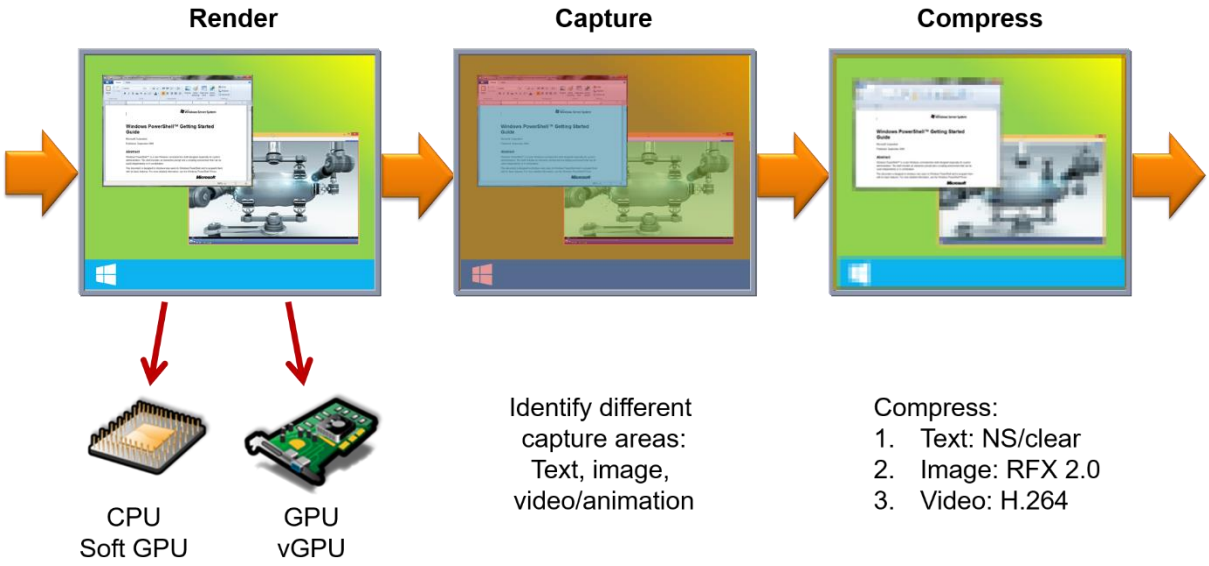
Maximale Resolution	Maximum number of monitors in virtual machine settings							
	1 Mon.	2 Mon.	3 Mon.	4 Mon.	5 Mon.	6 Mon.	7 Mon.	8 Mon.
1024 x 768	48 MB	51 MB	54 MB	57 MB	60 MB	63 MB	66 MB	69 MB
1280 x 1024	80 MB	85 MB	90 MB	95 MB	100 MB	105 MB	110 MB	115 MB
1600 x 1200	118 MB	125 MB	132 MB	140 MB	N/A	N/A	N/A	N/A
1920 x 1200	141 MB	150 MB	159 MB	167 MB	N/A	N/A	N/A	N/A
1560 x 1600	250 MB	266 MB	N/A	N/A	N/A	N/A	N/A	N/A

Starting with Windows Server 2012 R2, the default resolution that virtual machines with RemoteFX virtual GPU support is 1920 x 1200. This prevents manually changing the default resolution for full-screen sessions in most cases. In addition, RemoteFX virtual GPU also supports a higher maximum video RAM allocation. The maximum amount of VRAM for a Windows 8.1 VM on Windows Server 2012 R2 RDVH is 256MB dedicated VRAM plus 1GB shared VRAM. But only virtual machines with a minimum of 2,512MB of system memory will get the maximum of 1GB shared VRAM, which results in a total of 1,280MB dedicated + shared VRAM per VM. The VRAM configuration can be observed by running the command Dxdiag within a virtual machine's command prompt and clicking one of the Display tabs to view the Approx. Total Memory value.

The next major step forward in the RemoteFX vGPU evolution will be the upcoming Windows Server 2016 which is currently in the technical preview phase. Like in previous RDS versions, all major GPU vendors will be supported (NVIDIA, AMD and Intel). According to Microsoft, the new RemoteFX protocol features include OpenGL (up to 4.4), OpenCL (up to 1.1) and DirectX 12 support, 1GB dedicated VRAM configurable and decoupled from resolution, 4k screen resolution and H.264/AVC hardware offload support on clients. In addition, Windows Server 2016 RDS will support server VMs (= GPU-accelerated RDSH VM) and will come with a new RDSH desktop experience pack for improved usability and compatibility to the Windows 10 desktop. Microsoft also announced Hyper-V generation 2 VM support and pen remoting for tablets, such as Surface Pro 3. The RemoteFX performance will be improved with higher frames-per-second rates and lower jitter. A new implementation of Remote Desktop Gateway sockets simplifies load-balancing, with no IP affinity required.

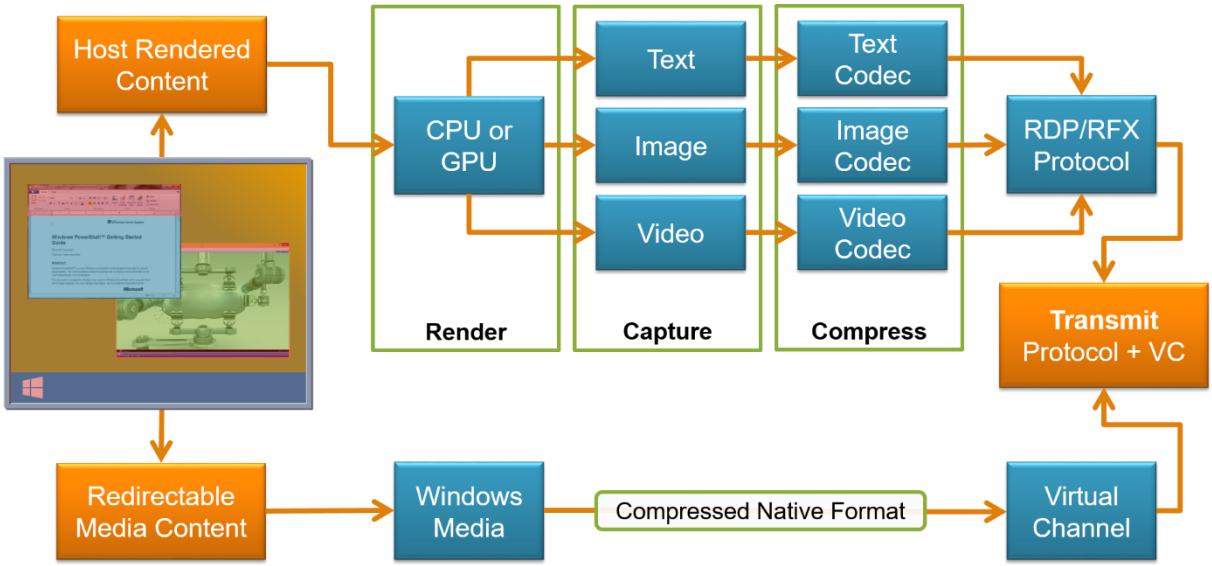
RemoteFX under the Covers

When Microsoft introduced RemoteFX version 1 in Windows Server 2008 R2 SP1, the only supported scenario was VDI (RDVH) in combination with a compatible physical GPU installed in the Hyper-V server. Starting with Windows Server 2012 this was changed in such a way that RemoteFX can be used with or without physical GPU. When no GPU is present, RemoteFX version 2 uses a CPU-emulated GPU (called SoftGPU) for rendering graphics. When, however, one or more GPUs are present in the server, the synthetic Hyper-V graphics driver in the VM intercepts and forwards a defined subset of graphics API calls to the physical GPU. This situation is referred to as RemoteFX vGPU - which should not be confused with Citrix XenServer vGPU. The advantage of the RemoteFX v2 architecture is that independent of the existence of a physical GPU in the server the same algorithms can be used for remoting the screen buffer content. Requirements for Windows Server 2012 (R2) RemoteFX vGPU are that all VMs must be running on Hyper-V, the server CPU must have Second Level Address Translation (SLAT) enabled and the graphics card(s) installed on the Hyper-V server must be DirectX 11 capable. Examples for RemoteFX-compatible GPUs are NVIDIA GRID K1 and K2 as well as AMD FirePro S10000, S9000 and S7000. A compatible GPU which was virtualized and shared by Hyper-V is available as an assigned GPU resource in the VM, almost like a physical PC with a physical GPU.



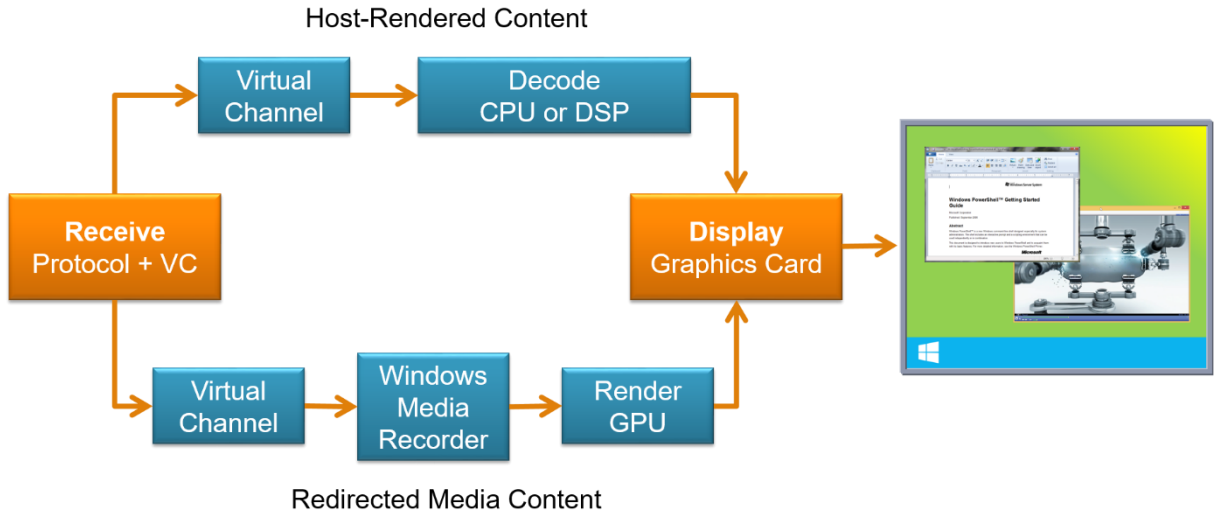
But RemoteFX v2 is not only about rendering, it's also about capturing and compressing screen frames. The RemoteFX capture component analyzes what was rendered in the previous frame and what's been rendered in the current frame. It dynamically divides the screen content into three different capture areas, representing text, still images and video/animation. Microsoft calls this mechanism Intelligent Screen Capture, which is responsible for checking screen content changes in the three capture areas and forwarding the changed bits for encoding. In addition, the capture component continuously communicates with the remoting protocol stack, allowing it to track network speed and understand which buffers and frames have been acknowledged by the client. As a result, the transmission of screen content can be adjusted dynamically according to the available bandwidth.

The final component is the RemoteFX encoder which is responsible for compressing the screen data before it is sent to the network protocol stack. Encoding can be done on the CPU, on the GPU or on dedicated hardware. In other words, the compress cycle may use both the GPU and CPU. Compression of the text area uses an NS/clear algorithm which provides high compression ratio and high quality without any noticeable time delay. Image compression uses either a RemoteFX-specific algorithm or AVC/H.264 in an adaptive manner, meaning that the image output quality continually improves within a couple of seconds. Video and animation content that is not redirected also uses H.264 for re-encoding. From the capture area encoder output the compressed RemoteFX content is passed on to the RDP stack where it will go through the "standard" processing, including bulk compression, encryption and prioritization. As much as possible, compression is offloaded to the GPU.



Typically, the entire screen goes through the RemoteFX render-capture-compress pipeline at 30 frames per second, except for redirected multimedia, such as WMV streams. Redirected media content is directly transmitted to the client in its compressed native format using an RDP virtual channel.

On the client side, the Remote Desktop Connection application receives the screen data from the remoting protocol network stack. It decodes the RemoteFX data using CPU, GPU or Digital Signal Processor (DSP), re-composes the entire desktop and displays the received frames on the local screen. Only redirected media content is processed in a data stream separated from the RemoteFX receiver pipeline. It is sent directly to the associated media player, typically supported by the GPU.



Vendor support for different 3D graphics solutions	Software 3D-Graphics	Bare Metal Client OS	Pass-Through GPU	GPU Virtualization GRID vGPs	Shared GPU for VDI (API Intercept)	Sharing GPU for XenApp/RDSH
Microsoft RDSH – Server OS (RDSH)	√	X	X	X	N/A	√**
Microsoft RDVH – Client OS (VDI)	√	X	X	X	√	N/A

NICE DCV

NICE vision is to provide high performance access for the most demanding applications, including high end 3D and computing tools, from virtually any kind of device and network. NICE have been pioneers in Technical Cloud deployments at leading companies and institutions, in many vertical markets including Oil and Gas, Industrial Manufacturing, Life Sciences, Government, Electronic Design, Financial Services and Research / Academic.

NICE is a software and consulting Services Company that is dedicated to improving ease of use and productivity, and achieving major hardware cost savings for companies running high performance technical applications. NICE developed the world's first HPC web portal, EnginFrame, some 15 years ago and this is widely used by engineers, geologists and others to ease the process of submitting HPC computations. EnginFrame is application aware, has a rich set of data management features and provides feedback to the user while the job is running. More recently, NICE acquired DCV from IBM, after many years of development, and has since further developed this software to the point where is it has a broader set of capabilities than any other software of its kind. NICE Desktop Cloud Visualization (DCV) enables remote visualization, and workstation virtualization for a mix of Windows and Linux HPC applications, all sharing graphics cards (GPUs).

Solution Details

NICE remote access software for 3D and HPC applications is built on two pillars: the EnginFrame portal and the Desktop Cloud Visualization (DCV) protocol.

EnginFrame provides a flexible Web-based framework to publish interactive, batch and full desktop services, a comprehensive set of data transfer and management features, coupled with a proven integration with mainstream HPC management software. It offers a user-friendly experience for any Private and Public Cloud computing need, from simple remote desktop brokering to the most advanced engineering workflows.

DCV delivers the highly responsive and network optimized experience needed for remote desktop and remote GUI access for high end applications based on OpenGL or DirectX. With its support for the latest GPU virtualization technologies from NVIDIA, real-time collaboration capabilities, hypervisor independence and compatibility with Linux and Windows OS, it allows users to be enjoy running high end 3D applications from any device and over most network conditions.

Providing both high-performance computing (HPC) and visualization tools is becoming more difficult than ever. The rising cost of workstation management, larger file sizes, more remote workers, and

the resulting need for collaboration are requiring IT professionals to reexamine how they deliver services.

NICE EnginFrame, NICE's HPC portal offering, is an advanced portal that provides access to grid-enabled infrastructures, HPC clusters, data, licenses, and interactive applications. It can be accessed by any authorized user with a standard web browser. EnginFrame handles computationally-intensive and sometimes parallel job submission, control, and monitoring. EnginFrame is based on standard protocols that facilitate the deployment of engineer-friendly portals to create, discover, and explore more efficiently. It provides for encrypted access and file transfers, protecting intellectual property and infrastructures.

EnginFrame provides HPC capabilities for large calculations, while Desktop Cloud Visualization provides the 3D modeling tools needed for complex, detailed models. Used together, they meet the computing needs for technical end users while delivering services in an intuitive, user-friendly interface.

Architecture

The NICE DCV architecture consist of:

- DCV Server, equipped with one or more GPUs, used for OpenGL rendering
- One or more DCV End Stations, running on "thin clients", only used for visualization
- Heterogeneous networking infrastructures (like LAN, WAN and VPN), optimized balancing quality vs frame rate

3D application delivery requires servers / blades equipped with NVIDIA GPUs, including many Quadro and Tesla models but best performing with the latest GRID cards, like the K2 or the gaming cards provided by Amazon G2 instances. A free accelerated endpoint software is provided for Linux, Windows and Mac OS/X, running on a broad range of devices including many thin clients or tablets, and other clients are also available on iOS, Android and other devices via our VNC compatibility mode. Thanks to its adaptive compression, DCV can accommodate high frame rates over LAN and WAN connections even on very high trans-oceanic latency and low bandwidth scenarios.

Besides a Web server layer for the portal publishing, the HPC application delivery leverages existing HPC clusters running leading job schedulers (including Torque/Moab, LSF, GridEngine, PBS/Pro, etc.) and mainstream batch applications, and requires only a Web browser to enjoy the full HPC Cloud experience.

The NICE DCV Technical Cloud software consists of three main components:

1. Self-service portal: The self-service portal enables engineers and scientists to access the applications and data in a web browser–based setting. It also provides security, monitoring, and management to ensure that users cannot leak company data and that IT managers can track usage. Engineers and scientists access applications and data directly from their web browsers, with no need for a separate software installation on their local client
2. Resource control and abstraction layer: The resource control and abstraction layer lies underneath the portal, not visible to end users. It handles job scheduling, remote visualization, resource provisioning, interactive workloads, and distributed data management without detracting from the user experience. This layer translates the user request from the browser and facilitates the delivery of resources needed to complete the visualization or HPC

tasks. This layer has a scalable architecture to work on a single cluster or server, as well as a multi-site WAN implementation

3. Computational and storage resources: The technical cloud's horsepower includes the company's existing or newly provisioned industry-standard resources, such as servers, HPC schedulers, memory, graphical processing units (GPUs), and visualization servers, as well as the required storage to host application binaries, models and intermediate results. These are all accessed through the web-based portal via the resource control and abstraction layer and are provisioned according to the end user's needs by the middle layer

Technical cloud software is built on common technology standards. The software adapts to network infrastructures so that an enterprise can create its own secure engineering cloud without major network upgrades. The software also secures data, removing the need to transfer it and stage it on the workstation, since both technical applications and data stay in the cloud. These solutions feature the best characteristics of cloud computing—simple, self-service, dynamic, and scalable, while still being powerful enough to provide 3D visualization as well as HPC capabilities to end users, regardless of their location.

The technical cloud software solution can be specialized and optimized in two primary ways:

- HPC portal
- Remote 3D "Virtual Workstations"

Remote 3D "Virtual Workstations"

This model is ideal for 3D visualization and connecting remote users to OpenGL applications that run in a data center. Users have access to full desktops - like in a common VDI (Virtual Desktop Infrastructure) – or individual applications and data through a web browser, but the 3D performance is much better than a traditional VDI. The "Virtual Workstation" model of technical cloud software primarily provides visualization capabilities. It includes visualization servers equipped with one or more GPUs that may or may not be shared among users. These GPUs provide OpenGL acceleration required by technical applications that can be installed and run in the cloud without modifications. The end user works from a thin client, such as a laptop equipped with a web browser, that is only used to show pixels and requires no data or application logic locally. Therefore, even tablets can be used to work or collaborate.

Data and applications are kept in the data center, but remote desktop access allows engineers and scientists to use them and be productive from remote locations. Technical cloud software can be designed and installed by in-house IT staff; however, technical cloud consulting firms can provide support throughout the implementation processes. Thus managers and architects can focus on delivering the required HPC and visualization capabilities while minimizing disruption to business operations.

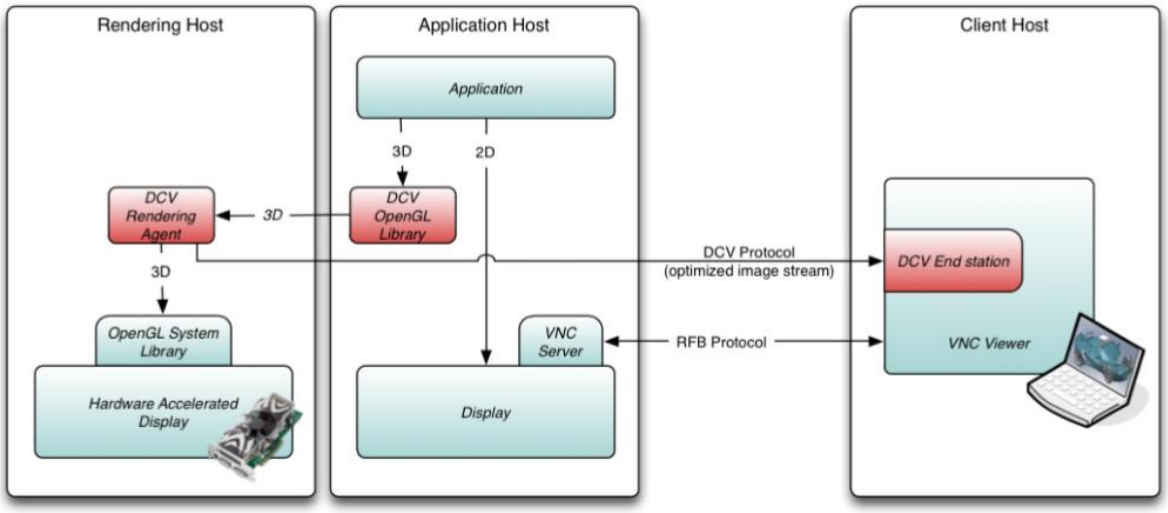


Image source: NICE

The OpenGL application runs in the Application Host, normally a Virtual Machine. All the OpenGL calls performed by the application are intercepted by the NICE DCV OpenGL interposing library and are redirected to an external rendering host.

The external rendering host must run the Linux OS and must be equipped with one or more (NVIDIA) 3D accelerated graphic adapters. For maximum performances, the advice is to use the application host and rendering host on the same physical host. The rendering host could be a Linux box running KVM hosting multiple Windows VMs acting as application hosts, or a Linux VM with GPU pass-through running on the same physical host as one or more Windows VMs using Citrix XenServer or VMware ESX.

In addition to sharing a single GPU among multiple Windows VMs, OpenGL applications can run on the Linux rendering host and you can share a GPU among multiple OSes, applications running in the Windows VMs and applications running in the Linux host.

Another technology NICE developed is the ability to share the same GPU among multiple users on the same physical Linux host, or Linux VM with GPU pass-through. In this scenario multiple users are all running on the same Linux instance leveraging the multi-user capability of Linux each one with his own desktop instance with full hardware accelerated 3D capabilities.

NICE DCV Highlights

The following list includes the NICE DCV highlights:

- Enables high performance remote access to interactive 2D/3D software applications on low bandwidth/high latency
- Supports multiple heterogeneous OS (Windows, Linux)
- Enables GPU sharing
- Supports 3D acceleration for OpenGL applications running on Virtual Machines
- Supports multiple user collaboration via session sharing
- Enables attractive Return-On-Investment through resource sharing and consolidation to data centers (GPU, memory, CPU, ...)
- Keeps the data secure in the data center, reducing data load and save time
- Enables right sizing of system allocation based on user's dynamic needs

- Facilitates application deployment: all applications, updates and patches are instantly available to everyone, without any changes to original code

NICE supports both Windows and Linux with Oracle Linux, SUSE Linux Enterprise Server and RedHat Enterprise Linux guest OS instances.

Both DCV and EnginFrame are licensed on concurrent usage: DCV based on the number of desktop sessions running on the servers, EnginFrame on the number of browsers connected to the portal. Licenses can be annual subscriptions or permanent with annual maintenance contracts to get access to the new releases and technical support.

Vendor support for different 3D graphics solutions	Software 3D-Graphics	Bare Metal Client OS	Pass-Trough GPU	GPU Virtualization GRID vGPs	Shared GPU for VDI (API Intercept)	Sharing GPU for XenApp/RDSH
NICE DCV	X	√	√	√	√	N/A

OTOY

OTOY is a Los Angeles-based cloud rendering company. Their aim is to deliver real-time, cinema-quality 3D graphics to filmmakers, 3D designers, animators, game developers, and end users by using the rendering power of the cloud, consisting of clusters of multiple GPUs, and streaming it to the client through a web browser.

OTOY launched the ORBX and OctaneCloud Amazon Machine Images (AMIs) on the AWS Marketplace in early November 2013. The use cases have ranged from HTML5 cloud gaming to high-end, plugin-free application streaming. OTOY wants to provide a dependable, royalty-free platform for delivering next gen video, cloud gaming and app streaming to any HTML5 browser that will make the open web the platform of choice for content creators and developers worldwide. Run any app in a browser -- no plugins required.

Solution Details

The OTOY infrastructure consists of two central components: ORBX.js and Octane Cloud Workstation

ORBX.js is a pure JavaScript framework developed in partnership with Mozilla and Autodesk to make the de-facto standard for consuming high-performance cloud graphics. ORBX.js enables state of the art 1080p60 cloud streaming to all HTML5 browsers, without the use of plug-ins, browser-specific video codecs, thin client-installations or native code dependencies.

The HTML client includes applet to launch native client from JavaScript (on PC/OSX/Linux/iOS) and run the session as a secure full screen or windowed process with much stronger encryption than a browser based https/wss stream. There is a beta HTML5 client for mobile available for Safari iOS, FireFox OS, Android Browsers (Chrome, Opera, Firefox, Maxthon). There is also a HTML 5 desktop beta client for Windows available.

ORBX enabled AMIs on Amazon EC2 provide a complete Windows or Linux desktop or workstation replacement in the cloud, capable of installing and running software or services designed for

traditional PCs. This means that PC software and games can be easily whitelisted to run remotely from Amazon EC2 without modification, including day and date deployment of games and apps through digital delivery platforms such as Valve's Steam OS or Ubisoft's Uplay.

Octane Cloud Workstation is a turn-key high performance cloud desktop solution specifically designed for streaming high-end remote graphics. Use this AMI to stream a Windows based virtual desktop, hosted in the cloud, to a web browser anywhere in the world. - Access your performance intensive Windows-based applications and data from any device, regardless of your operating system or device performance. - Your applications seamlessly render with clarity in HD due to OTOY's next generation ORBX Video Codec.

Stream your applications using OTOY's native client application (available for Windows, Linux and iOS), or go plugin-free with ORBX.js on any modern browser. The solution includes OTOY WebCL remote graphics driver, the only OpenCL 1.2 GPU runtime for NVIDIA GRID.

Highlights:

- ORBX.js delivers 60 Hz 'zero client' HD cloud desktops to any HTML5 web browser using pure JavaScript. No plugins required!
- Access to a fully licensed version of Octane Render Cloud Edition - enabling real time GPU rendering on EC2
- Supports GPU graphics applications (DirectX 9/10/11, OpenGL 4.x, CUDA, OpenCL) on G2 instances. OTOY WDDM Aero virtual desktop resolution up to 2048 x 1536. Validated to work with high performance 2D graphics and media applications on all supported instance types

Host GPU driver support:

- ORBX GPU Desktop Graphics: OTOY WDDM Aero virtual desktop (up to 2048x1536), DirectX 11, OpenGL 4.x, DXGI 1.1+ Application support
- ORBX Render: Built in OTOY Octane Render with integration across all major ADSK apps: 3DS Max, Maya, XSI, AutoCAD, Revit, Inventor
- ORBX Compute: Built in OTOY OpenCL 1.2 driver/runtime/compiler for NVIDIA Kepler. Built on top of NVIDIA CUDA driver, compiled OpenCL directly to PTX, used for encoding/rendering (3rd party SDK being developed in partnership w/ NVIDIA)

Virtual Machine specifications

- Base OS: Windows, Windows Server 2008 R2 2008R2 x64
- Memory 15 GB
- CPU 22 EC2 Compute Units (8 virtual cores), plus 1 NVIDIA GK104 GPU
- Storage 1 x 60 GB SSD
- Amazon EC2 GPU G2.2XL instance

Host Device I/O support:

- Remote audio driver: speaker output, microphone input
- Remote USB HID device driver (requires clients that support via raw input, could be extended to support 3D Mouse, Wacom tablet, Oculus Rift, Leap Motion, etc.)
- Printer driver: supports remote image based printing and large format printer support
- Xinput driver: analog, gamepad, battery level, rumble pack support
- Clipboard driver: text, images, metafiles, OLE objects used by AutoCAD, Microsoft Office, etc.

- Droptarget process monitor: Drag and drop file and folder support (bi-directional on desktop clients)
- Webcam driver (client-to-host, not available for all clients, in beta)

VMware Horizon

VMware was founded in 1998 and in the early years focused on providing a platform to run Windows- and Linux-based machines virtually. In 1999 this resulted in VMware Workstation. After the launch of ESX in 2001, VMware made a name in server virtualization and quickly became the market leader in this segment.

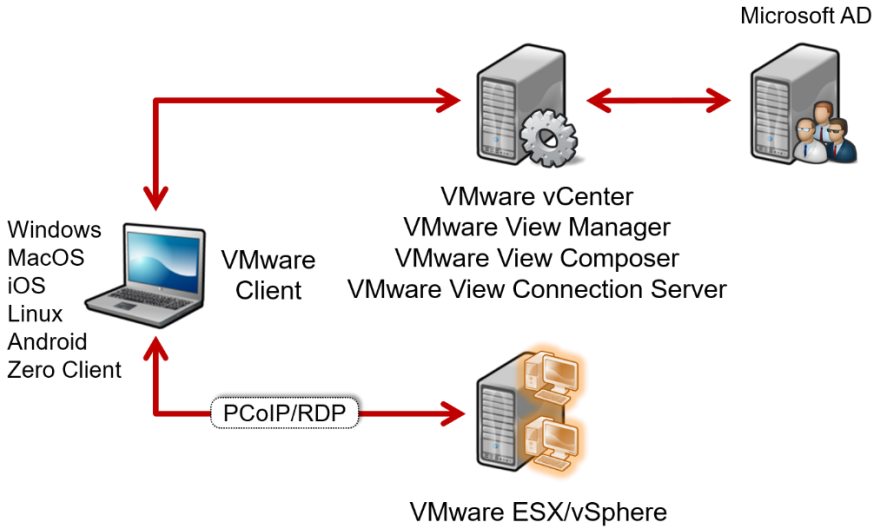
In May 2007, VMware acquired UK-based ISV Propero. In early 2008 VMware used some of the Propero software assets to enter the VDI market with the launch of VMware VDM 2.0. VMware VDM gave users a centrally hosted desktop. The name of VDM evolved into a new product suite: VMware View. In 2009 VMware partnered with Teradici for their display protocol PC-over-IP (PCoIP) which was specifically developed for server hosted (virtual) desktops and Blade PCs. This partnership allowed VMware to make a change to the remote display capacity that resulted in a major enhancement to the user experience. PCoIP technology is delivered in both hardware and software implementations. VMware View 4.x and above are based on the software version of PCoIP.

It is important to note that the PCoIP remoting protocol is only responsible for the graphics portion of remote desktops. Other functionalities, such as multimedia redirection, Unified Communications, Webcam redirection and USB mapping, application remoting were added by the VMware Horizon team. With VMware View 5.3, VMware introduced the Blast protocol for HTML Access.

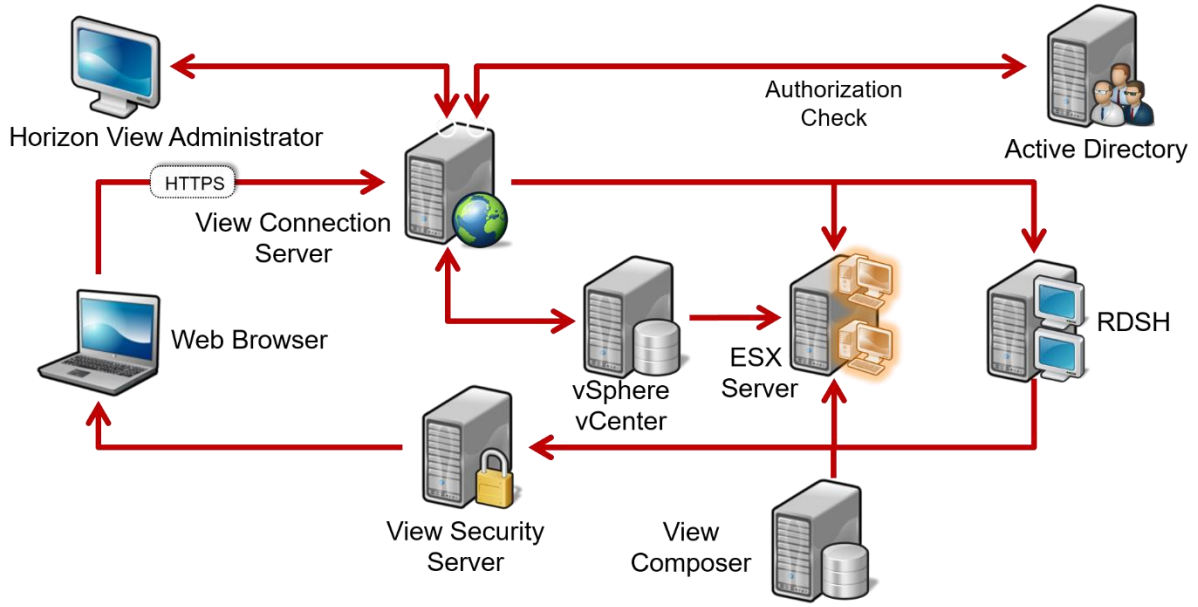
VMware recently introduced availability of Horizon for Linux. With this solution, organizations can deliver, manage and access Linux virtual desktops and eliminate the need to license commercial operating systems. Horizon for Linux desktops provides single sign on access to office, 3D graphics (vGPU and vDGA) and developer applications. Horizon 6 for Linux is ideal for software developers, CAD/CAM developers, government workers, cost sensitive deployments and organizations who want to take advantage of the cost savings, security and customizations available with Linux.

Architecture

The VMware Horizon architecture is shown below. Apart from the VMware Virtual infrastructure, you only need the View Connection Server to be up and running. To use the linked-clones technology, the VMware View Composer is installed on the VMware vCenter server. With this feature it is possible to use one (snapshot of a) golden image to deploy virtual desktop VMs and save disk space because the VMs use the same golden image and an additional Delta file of changes. The protocols used with VMware View/Horizon are RDP and PCoIP.



With the combined additions of App Volumes, Identity Manager, and User Environment Manager (UEM) in Horizon 6, VMware now offers Workspace Environment Management (WEM). Organizations now benefit from a real-time app delivery and lifecycle management platform, with integrated persona management and customization capability layered into the user workspace dynamically, based on role, complete with a unified workspace for all desktops and apps. With Horizon 6, organizations can now effectively leverage stateless, non-persistent desktops delivered to end users that feel and behave like persistent – delivering a user experience akin to their trusted physical desktop predecessor.



Licensing

VMware Horizon is offered in three editions, aligned with capabilities that organizations want to leverage in their desktop/app virtualization environment. See the table below to compare Horizon Standard, Advanced and Enterprise Editions.

Horizon Edition:		Linux	View Standard	Advanced	Enterprise
Packaging quantity		10-pack, 100-pack			
Licensing Options NU= Named User, CCU= Concurrent Connection User		CCU	CCU	NU or CCU	NU or CCU
Desktop Infrastructure					
Virtual Desktops	VMware Horizon with View		x	x	x
	VMware Horizon for Linux	x			x
Cloud Infrastructure	VMware vSphere and vCenter for Desktop	x	x	x	x
Image Management for Virtual Desktops	VMware Mirage and Fusion Pro			x	x
Applications					
Application Catalogue and Dashboard	XA, RDSH, SaaS, ThinApp, all accessible through VMware Identity Manager Std Ed.			x	x
Application Packaging	VMware ThinApp		x	x	x
Application Virtualization	Apps Publishing (RDSH)			x	x
Just-in-time App Delivery	VMware App Volumes				x
Personalization and Policy					
Persona Management	VMware User Environment Manager				x
Storage					
Virtual Storage	VMware vSAN for Horizon, Advanced Edition			x	x
Operations Management					
Operations Dashboard- Health Monitoring & Performance Analytics	vRealize Operations for Horizon				x
Capacity Management – Planning & Optimization	vRealize Operations for Horizon				x
Cloud Automation					
Design & automate workflows	vCenter Orchestrator + Desktop Plugin				x

Image source: VMware

3D Graphics for Virtual Desktops with VMware Horizon

VMware offers a complete spectrum of capabilities for delivering rich, immersive 3D graphics aligned with use case and required level of performance, ranging from basic Task Workers to High-End Engineers and Designers. The software-based approaches for 3D graphics are referred to as “SVGA”, “Soft3D” and “VMware SVGA 3D”. These are based on a VMware Windows Display Driver installed in the Windows+ guest OS. For higher-end use cases requiring workstation-equivalent performance, VMware offers a Pass-through GPU technology called VMware vDGA, Virtual Dedicated Graphics Acceleration. With this approach, GPU access is provided by a native NVIDIA driver accessible to 3D applications within the guest OS with a 1:1 mapping between virtual desktop and the NVIDIA GRID GPU.

VMware Horizon supports two modes of GPU sharing: These are VMware vSGA (Virtual Shared Graphics Acceleration) and vGPU. With vSGA, multiple virtual machines share the physical GPUs within the NVIDIA GRID card installed in the ESXi host, providing hardware accelerated 3D graphics to virtual machines. In each virtual machine a VMware display driver is installed that provides emulated access, mediated by vSphere, to the underlying GPU hardware. Applications designed or certified for

NVIDIA or AMD GPUs need to be recertified against the VMware driver since no native NVIDIA driver is presented. vSGA offers some capabilities some environments may require including support for vSphere vMotion and High-Availability features. vSGA is limited however, to DirectX 9.0c and OpenGL 2.1. With vGPU, VMware and NVIDIA provide a highly evolved, next-generation platform for high-performance 3D graphics supporting shared GPU environments. vGPU employs an NVIDIA GRID Manager resident within vSphere to multiplex graphics calls made to the GPU from multiple guest VM's to the underlying GRID GPU. This approach essentially bypasses the hypervisor, reducing latency and increasing graphics rendering performance. Each guest VM leverages a native NVIDIA driver for maximum application compatibility.

VMware's hardware-accelerated graphics for Desktop Virtualization timeline is as follows:

- **2008**: VMware supports Teradici with view Software/Hardware1:1
- **2009**: Availability of PCI Pass-through
- **2011**: VMware View 5.0 support for soft3D with CPU rendering
- **2012**:
    - Support migration of VM's with vSGA.
    - vDGA in vSphere 5.1 and introduction to vSGA in Horizon View 5.1
- **2013**:
    - vSGA vSphere 5.5 support for AMD GPUs,
    - vSphere 5.5 support for vDGA,
    - GPU Pass-through support for Horizon View 5.3 and Citrix XenDesktop/XenApp.
    - VMware Horizon View 5.3+ supports vDGA with all versions of OpenGL and DirectX
- **2014**:
    - VMware announced vGPU support for VMware vSphere and Horizon View at NVIDIA GTC 2014. Release date 2015
- **2015**
    - NVIDIA vGPU support with vSphere 6.0+ and VMware Horizon 6.1+
    - NVIDIA GRID 2.0 support with vSphere 6.0+
    - Horizon for Linux supports 3D with vDGA with NVIDIA GRID with Horizon 6.1.1+
    - Horizon for Linux supports vSGA with Horizon 6.2+
    - Horizon for Linux supports shared 3D with NVIDIA GRID vGPU 2.0 with Horizon 6.2+
    - RDSH & Application Remoting with 3D supported with Horizon 6.2+

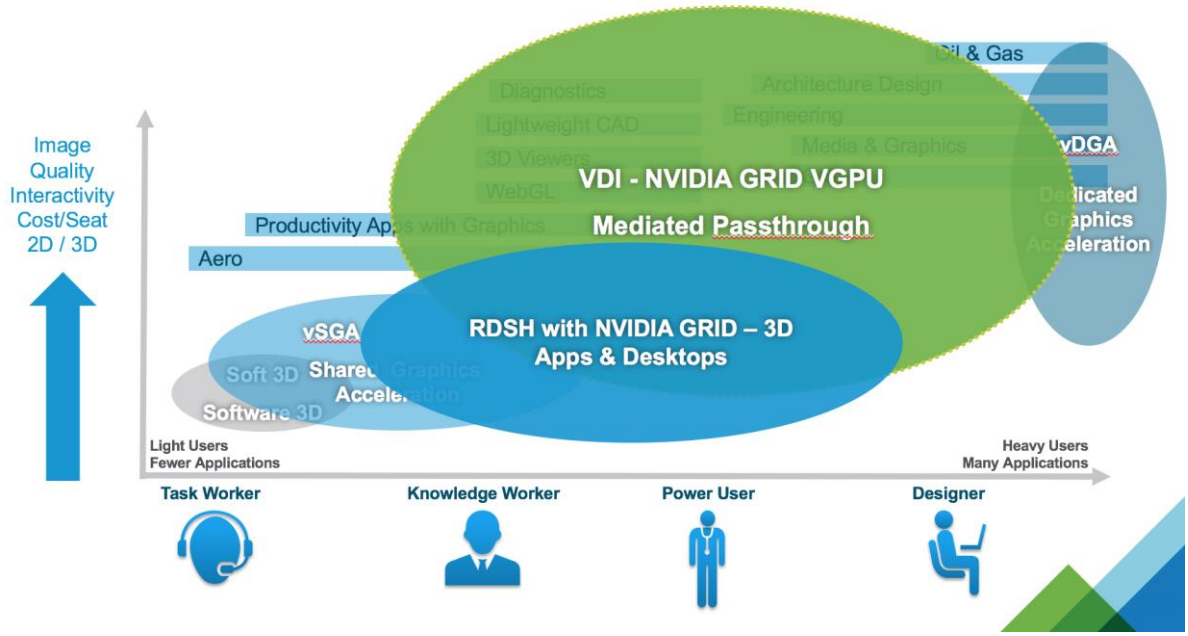


Image source: VMware

Vendor support for different 3D graphics solutions	Software 3D-Graphics	Bare Metal Client OS	Pass-Trough GPU	GPU Virtualization GRID vGPs	Shared GPU for VDI (API Intercept)	Sharing GPU for XenApp/RDSH
VMware Horizon	√	√	√	√	√	N/A

3D Graphics for Virtual Desktops Vendor Solutions at a Glance

The diagram below gives a complete overview of the 3D graphics for virtual desktops solutions.

Vendor support for different 3D graphics solutions	Software 3D-Graphics*	Bare Metal Client OS	Pass-Trough GPU	GPU Virtualization GRID vGPs	Shared GPU for VDI (API Intercept)	Sharing GPU for RDSH**
Citrix XenApp – Server OS (RDSH)	√	N/A	√	√	N/A	√
Citrix XenDesktop – Client OS (VDI)	√	√	√	√	X	N/A
NICE DCV	X	√	√	√	√	N/A
Microsoft RDSH – Server OS (RDSH)	√	X	X	X	N/A	√**
Microsoft RDVH – Client OS (VDI)	√	X	X	X	√	N/A
VMware Horizon	√	√	√	√	√	√

*) Level of capabilities dependent on OS and 3D APIs

**) Only Bare Metal GPU on RDSH brings 3D API support. No GPU Pass-through in Hyper-V Virtual Machine scenario. Rendering content leverages GPU capabilities. RDP/RemoteFX Capture and

Compress cannot take advantage of GPU. Bare Metal only required for HyperV, XenServer and vSphere both support pass-through and vGPU for RDSH options.

The following table compares the support of common graphics formats for all products introduced in this chapter.

	XD7.6 VDA SoftGPU	XenApp Shared GPU	XD7 HDX 3D Pro Pass-Thru GPU	XD7.6 GRID vGPU	RDVH Win7/10 RemoteFX	RDVH Win10 RemoteFX	Horizon 6.2 vDGA Pass-Thru GPU	Horizon 6.2 vSGA Shared GPU	Horizon 6.2 GRID vGPU
DirectX 9	√	√	√	√	√	√	√	√	√
DirectX 10	X	√	√	√	√	√	√	X	√
DirectX 11	X	√	√	√	√	√	√	X	√
DirectX 12	X	√	√	√	√	√	√	X	√
OpenGL 1.1	√	√	√	√	√	√	√	√	√
OpenGL 2.0	√	√	√	√	X	√	√	√	√
OpenGL 3.0	X	√	√	√	X	√	√	X	√
OpenGL 3.3	X	√	√	√	X	√	√	X	√
OpenGL 4.0	X	√	√	√	X	√	√	X	√
OpenGL 4.4	X	√	√	√	X	√	√	X	√
OpenGL 4.5	X	√	√	√	X	√	√	X	√
OpenCL 1.1	X	√	√	X	X	√	√	X	X
OpenCL 2.0	X	√	√	X	X	√	√	X	X
OpenCL 2.1	X	√	√	X	X	√	√	X	X

Enabling Technologies – GPUs and CPUs

AMD

Founded in 1969, AMD designs and integrate technology that powers intelligent devices. AMD acquired ATI Technologies Inc in 2006. ATI was a semiconductor technology corporation based in Markham, Ontario, Canada, that specialized in the development of graphics processing units and chipsets. The acquisition of ATI in 2006 was important to AMD's continual development of GPU related technology and system technology like HSA (Heterogeneous System Architecture). Since 2010, the ATI brand name is no longer used for any of AMD's graphics processor products. The branded GPU product ranges are Radeon for consumer and Firepro for professional graphics. AMD combined system (CPU+GPU) technology drives the leading gaming console solutions. AMD believes that providing pixel rendering and compute via GPU are essential for delivering high quality 3D, multimedia and high performance compute experiences on next generation cloud based services.

AMD SKY delivers graphics rendering solutions for:

- Enterprises (on premises VDI and remote graphics for application or desktop delivery)
- Cloud Service Providers (gaming, consumer, DaaS and 3D/pixel rendering dependent services)

AMD STREAM delivers solutions for compute, leveraging a vendor neutral industry standard programming environment called OpenCL. OpenCL is a primary feature of the latest Apple MacPro systems that utilized FirePro graphics. AMD supports the S series GPU boards for both SKY and STREAM use cases. The same boards for VDI can be used for compute.

Radeon and FirePro

ATI launched the Radeon line in 2000 as their consumer 3D accelerator add-in cards, its flagship product line and the direct competitor to NVIDIA's GeForce. Mobility Radeon is a series of power optimized versions of Radeon graphics chips for use in laptops.

AMD CrossFireX was ATI's response to NVIDIA's SLI platform. It allowed, by using a secondary video card and a dual PCI-E motherboard based on an ATI Crossfire-compatible chipset, the ability to combine the power of the two, three or four video cards to increase performance through a variety of different rendering options. There is an option for an additional PCI-E video card plugging into the third PCI-E slot for gaming physics, or another option to do physics on the second video card.

The FirePro product line was launched in 2001, following ATI's acquisition of FireGL Graphics from Diamond Multimedia. At the beginning, they were workstation CAD/CAM video cards, based on the Radeon series. Since that time the software and hardware development changed dramatically with latest generation FirePro products providing modern drivers with high stability and real world application performance in mind. According to the SPECviewperf 12 benchmark, AMD FirePro products are some of the higher performing solutions across the professional graphics market.

The Role of AMD in 3D Graphics for Virtual Desktops

AMD's ambition is to be compatible with the leading platforms in the VDI market. They are driving knowledge worker and high performance workstation virtualized solutions with VMware. AMD is engaged with Citrix on application and session delivery for remote users. And they are engaged with

the Microsoft RemoteFX/Hyper-V product teams, making AMD FirePro one of the best performing GPU for Microsoft RemoteFX.

- AMD FirePro for VMware Horizon View vSGA (shared graphics): One or more AMD FirePro S Series GPUs are attached to a server running VMware ESXi. The VMware graphics driver is used in each virtual machine, providing DirectX 9 and Software OpenGL. The ideal use cases for this setup are 3D Viewers/Model component viewing, GIS viewers, education/training environments and enterprise power users.
- AMD FirePro for VMware Horizon View vDGA (pass-through GPU): AMD FirePro S and W Series GPUs are directly mapped to individual virtual machines on VMware ESXi. The graphics driver in the virtual machines is provided by AMD to ensure full performance and workstation ISV support along with full OpenCL support. Ideal uses cases are MCAD/CAE/PLM midrange users, GIS power users, application classroom/training and media content creation.
- AMD FirePro for Microsoft RemoteFX on Hyper-V (shared graphics): One or more AMD FirePro S Series GPUs are attached to a server running Windows Server 2012 with Hyper-V role enabled. The driver on the Hyper-V platform is provided by AMD. The virtual machines have shared access to the GPU through the Microsoft synthetic driver. Ideal use cases are 3D viewers/model – component viewing (DirectX only), GIS viewers, education/training environments and enterprise power users.
- AMD FirePro for Citrix XenApp/XenDesktop: Prerequisite is a physical server with Microsoft Windows Server 2008 R2 or Windows Server 2012 R2 with AMD S Series GPU plus driver. The setup can be either Direct Connected GPU to Windows Server or Direct Pass-Through via Hypervisor, providing GPU support to Windows Server session running XenApp or XenDesktop session services. The AMD driver allows the delivery of graphics content with HDX 3D. Ideal use cases are remote workstation users, tablet delivery for factory floor CAD access and packaging of applications to users needing dataset/CAD viewing only.
- AMD FirePro R5000 Remote Workstation: The AMD FirePro W5000 class GPU is combined with a Teradici PCoIP hardware encoder for highest performance. The GPU and the Teradici encoder are used together for accelerating remote access.

	FirePro S10000	FirePro S9170	FirePro S9150	FirePro S9100	FirePro S9050	FirePro S7000	FirePro S4000X	FirePro R5000
GPU	2 x Tahiti	Grenada	Hawaii	Hawaii	Tahiti	Pitcairn		Pitcairn
GPU clock	825 MHz	930 MHz	900MHz	824 MHz	900MHz	950MHz		825MHz
Stream Processors / shading units	2x 1792	2816	2816	2560	1792	1280	640	768
Memory size and speed	6GB/GPU 12Gb total	32GB 320GB/s	16GB – 320GB/s	12GB - 320GB/s	12GB – 264GB/s	4GB - 154GB/s	2GB	2GB - 102GB/s
Memory interface	384-bits	512-bits	512-bits	512-bits	384-bits	256-bits		256-bits
Form factor	full height/ full length	full height/ full length	full height/ full length	full height/ full length	full height/ full length	full height/ full length	MXM	full height/ full length
Core clock speed								
Max. power	375W	275W	235W	225W	225W	150W	45W	150W
Peak double precision (TFLOPS)	1.48	2.62	2.53	2.11	.806	.152	.62	.792
Peak single precision (TFLOPS)	5.91	5.24	5.07	4.22	3.23	2.4	.992	1.3
Teradici processor	N/A	N/A	N/A	N/A	N/A	N/A	N/A	Tera 2240
Cooling	Passive heat sink	Passive heat sink	Passive heat sink	Passive heat sink	Passive heat sink	Passive heat sink	N/A	Actively cooled

Note: AMD boards supporting Multiuser GPU are not available yet (Oct. 2015)

Multiuser GPU

At VMworld 2015, AMD demonstrated a new hardware-based GPU virtualization solution that is still in development, named AMD Multiuser GPU. Rather than repurposing an existing GPU and adding a software layer to accommodate virtualization requirements, AMD's approach was to create an entirely new class of GPUs with virtualization capabilities built into the silicon. AMD challenged the notion that the goal of GPU virtualization requires a software solution.

AMD Multiuser GPU is built around the industry standard SR-IOV (Single Root I/O Virtualization) technology that allows a single PCIe device to appear to be multiple separate PCIe devices. So in essence, it provides a way for physical devices to expose hardware virtualization. The SR-IOV technology is not new, it is also used to assign physical network cards to virtual machines. The AMD implementation of SR-IOV for Multiuser GPU is baked into the physical graphics card and requires support in the server BIOS. The result is a virtualized workstation-class experience with full ISV certifications.

By incorporating the well-developed SR-IOV specification as a foundation for the technology, AMD accomplished their goal with a hardware virtualized GPU. The SR-IOV specification supports the notion of a device becoming a physical function (PF) on the PCIE bus. Through the specification, each PF has the ability to create instances of itself called virtual functions (VFs). These VFs inherit the attributes of the parent PF allowing each to become fully capable mirror images of the PF. The PF controls access by the VFs to internally mapped resources such as the graphics core. The task of virtualization is therefore charged to the PF, unburdening the hypervisor from complex scheduling routines. The hypervisor need only to understand that, once the Multiuser GPU technology is enabled, number of additional fully functional graphics devices will appear on the bus in place of the parent physical device. A simplistic driver programs the SR-IOV parameters within the physical function and alerts the hypervisor to the presence of the virtual functions. Once the hypervisor becomes aware of these new devices, it allows them to pass through to the various virtual machines.

The PF manages sharing of graphical resources by scheduling time slices across VFs and allocating dedicated local memory to each of these VFs. The PF also assigns internal register spaces to each VF ensuring an orderly and structured method for the VFs to access resources and data, at the same time keeping that data secure. Because each VF inherits the attributes of the PF, it sees full GPU resources during its scheduled time slice. When these VFs are passed through to their assigned virtual machines, they will appear as full-featured graphical devices to the virtual machine's guest OS. Since the guest OS sees the VFs as a native graphics device, AMD's native FirePro graphics driver (for professional graphics devices) can be loaded within virtual machine to unlock the GPU's rendering capabilities.

AMD Multiuser GPUs support 2 to 16 VFs. Mapping one VF to a virtual machine allows the creation of up to 16 independent desktops that originate from a single GPU. User density is limited only by the availability of PCIE slots. Platforms that can support four GPU devices has the potential of supporting up to 64 users with 64 independent environments. This technology has only minimal impact on the hypervisor as it does not require any software stack on the host side that goes beyond a fairly simple driver. This driver sees up to 16 virtual GPUs (VFs) per physical GPU, with device virtualization and GPU resource management purely implemented in hardware. It is important that the BIOS is capable of dealing with all these "new" virtual GPU devices that cannot be distinguished from physical GPUs. Among the most important requirements is 40-bit memory addressing as the memory range must support the reservation of memory for these additional GPU devices on the PCIe bus.

The Multiuser GPU technology does not support the concept of profiles. The host administrator is given complete freedom to configure the GPU for 2 to 16 users and all configurations in between. A five user configuration is just as valid as a seven, two or sixteen user configuration. This concept has a lot in common with GPU pass-through as each hardware-virtualized GPU has a one-on-one relationship to a virtual machine. In the current implementation demonstrated at VMworld, AMD Multiuser GPU supports VMware vSphere/ESXi 5.5/6.x and the remoting protocols coming with Horizon View, Citrix XenDesktop and Teradici Workstation Host Software. The native AMD driver installed in each guest operating system supports OpenGL, DirectX and OpenCL on the hardware-virtualized GPU. As long as the hypervisor supports SR-IOV and is aware that a GPU device is installed there are no additional requirements from the hypervisor to manage virtualization on this GPU. Because of this simplified approach to unburden the hypervisor from all tasks related to GPU virtualization, the AMD Multiuser solution becomes hypervisor agnostic. This means that future support of hypervisors other than VMware vSphere is also possible.

A key benefit of AMD's hardware-based GPU virtualization technology is that one user cannot affect another user's GPU performance as GPU compute resources and video RAM are exclusively assigned to individual VMs. In other words, GPU compute cycles are not shared, each cycle is "owned" by one particular VM. The sum of the virtual GPUs (VFs) directly relates to the physical GPU (PF). The performance of an individual VF is therefore inversely proportional to the number of VFs configured per GPU. The fixed time slice apportioned to each VF ensures that each VF receives its fair share of GPU cycles. Consequently, one virtual machine's usage of the GPU resources does not hinder another's ability to use the same resources; one user cannot affect another user's performance.

Another interesting aspect of the Multiuser GPU solution is its ability to preserve the data integrity of virtualized desktops. Because of the hardware implementation of the solution and the use of dedicated local memory, there is a natural isolation of data between the VFs (and therefore the VMs). One VM is unable to peer into the desktop data of another VM. With security being a bare minimum requirement for any virtualization solution, AMD's hardware-based virtualized GPU solution offers a strong deterrent to unauthorized users who traverse the software or application layers seeking means to extract or corrupt GPU user data from the virtual machines. Although a VF can access full GPU capabilities at its particular time slice, it does not have access to the dedicated local memory of its sibling VFs. It also does not have access to hardware level SR-IOV or GPU registers that can change the nature of the virtualization environment or provide statistics of the underlying physical GPU. This privileged register access is reserved for the PF located at the host. Because the Multiuser GPU driver on the host side simply sets up parameters for SR-IOV and virtualization in hardware, no information about the individual VF/VM's desktop context is passed around in this layer. All data and information about context, time slices, synchronization and scheduling is locked in the silicon layer.

Later this year, AMD is planning to release different single GPU and dual GPU cards that are the foundation of Multiuser GPU. Unfortunately, only newer server hardware will be compatible to the way AMD Multiuser GPU implements SR-IOV, which eliminates many existing standard server platforms from this technology. More details about compatible servers and required BIOS specifications will be published by AMD in the near future.

Intel

Intel Corporation, headquartered in Santa Clara, California, is one of the world's largest manufacturer of CPUs, mainboard chipsets, network cards, solid state drives and embedded processors. Since many years, Intel is dominating the PC microprocessor and the mobile PC microprocessor market.

Intel VT

Intel Virtualization Technology (Intel VT) provides hardware support that simplifies processor and platform virtualization. In particular, this helps enabling reductions in Virtual Machine Monitor (VMM) complexity and providing considerable improvements in performance over software-only VM hosting. VT is an "umbrella" term and the overall capabilities are often categorized as VT-x, which refers to processor capabilities. VT-d is relating to chipset capabilities and VT-c is referring to network and connectivity capabilities. It is important to note that VT-d and VT-c are not dependent on VT-x. For example, a VT-x enabled system can operate without VT-d enabled or configured. In such a case you are simply missing the benefits of the feature.

Intel's family of Xeon server processors provides support for hardware-based technologies enabling desktop and applications virtualization and security. For example, with VT-x the Virtual Machine Monitor can run in a highly privileged Root Mode. An extended page table (EPT) allows the VMM to efficiently and securely manage memory for each guest VM. I/O subsystem performance enhancements and a technology called Direct Device Attach (DDA), where a guest VM can have access to a particular I/O device, are also on the long list of VT capabilities.

An additional Intel technology called VMCS Shadowing reduces the frequency with which the guest Virtual Machine Monitors (VMM) must access the root VMM in a nested environment. This technology is used with new usage models that allow two or more VMMs to be hosted on the same client system, both physical and virtual. With Intel VMCS Shadowing, the root VMM is able to define a shadow VMCS in hardware. A guest VMM can access this shadow VMCS directly and without interrupting the root VMM. Due to the fact that Shadow VMCS is implemented in hardware its performance is almost as good as in a non-nested environment.

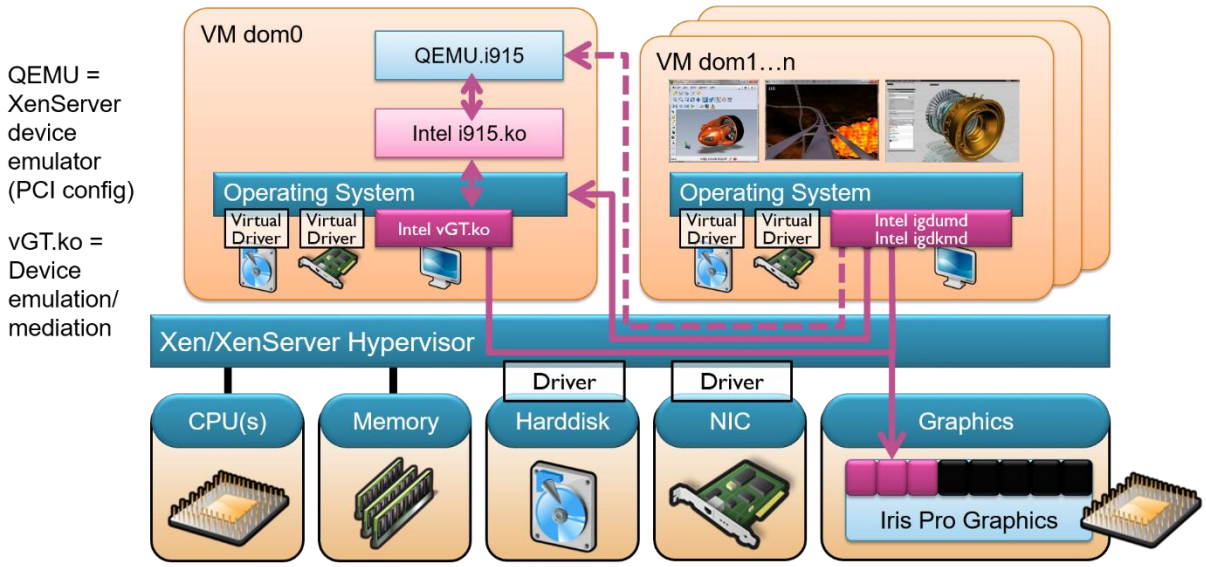
Security solutions such as Bromium and Intel DeepSAFE can leverage Intel VMCS for delivering secure solutions in a VDI environment for 3D graphics.

Intel GPU

Intel offers integrated processor graphics technology in their Xeon E3 line of CPUs. A variety of processor configurations are available based on the GT2 and GT3e graphics components. In September 2013, Intel announced their Haswell CPUs with four models of integrated GPUs. The high-end model is called GT3e or Iris Pro with 40 execution units, which is equivalent to 40 GPU cores. In addition to the 40 GPU cores, the Iris Pro model comes with 128 MB of embedded DRAM acting as a Level 4 cache. It is shared dynamically between the GPU and CPU, allowing for high data transfer rates.

Only three months after the Haswell announcement, Intel also announced XenGT (recently renamed to Intel Graphics Virtualization Technology or Intel GVT-g) as a GPU virtualization solution with mediated graphics pass-through. As the name indicates, Intel GVT-g combines Xen hypervisor and Iris Pro GPU cores (GT3e). A virtual GPU is assigned to a guest VM, with resource assignment brokered and managed by a "mediator" component running in the Dom0 session. Controlled by the mediator, Intel's native graphics driver runs inside each guest VM, communicating directly with the assigned

GPU cores and preventing hypervisor intervention in performance critical paths. On the hardware side, Haswell Iris Graphics is Intel's main focus regarding hardware-accelerated graphics remoting along with next-generation Broadwell processors. In September 2015 Citrix released a technology preview of GVT-g on XenServer, which was first demonstrated at Synergy back in Q2 2015.



In addition to GVT-g (mediated GPU pass-through), Intel has two more GVT technologies that can be used in conjunction with Iris Pro. GVT-d is commonly known as Direct Graphics Adaptor (vDGA) and it allows direct assignment of an entire GPU's power to a single user, passing the native driver capabilities through the hypervisor without any limitations. The assignment of the GPU is accomplished using Intel's foundational hardware virtualization features namely VT-d. Citrix worked with Intel to add GVT-d support to XenServer 6.5 SP1 which shipped in May. Intel GVT-s is better known as Virtual Shared Graphics Adaptor (VMWare vDGA) and RemoteFX vGPU (API Intercept). This graphics virtualization approach requires a synthetic graphics driver in the VM and uses an API forwarding technique to interface with the Intel graphics hardware.

In summer 2015, Intel started shipping the new generation "Broadwell" APU, which is an Intel Xeon E3-1200 v4 processor with Iris Pro Graphics P6300. This integrated GPU comes with 48 execution units (cores) with 128MB of eDRAM. It can deliver up to 1.8 times the 3D graphics performance of the previous generation Intel Xeon E3-1200 v3 with Intel HD graphics. The memory interface of the Xeon E3-1200 v4 Iris Pro graphics unit allows to address up to 32GB DDR3(L)-RAM used as vRAM. At VMworld 2015, Intel demonstrated this APU in combination with VMware vSphere and Horizon View, delivering GPU-accelerated user sessions. Citrix XenServer and XenDesktop can take advantage of the Intel APU in the same way.

The design of the Intel E3-1200 v4 with integrated Iris Pro graphics unit is geared towards low to medium-end remote graphics workstation use cases. It is a low-power and low-cost solution for customers who want to keep their critical data in centralized datacenters and deliver rich 3D applications over local and wide-area networks. According to Intel, their GVT solution allows Xeon E3-1200 v4 customers to dedicate the resources of each processor to a single designer or share them among multiple less demanding users. The power consumption of a Xeon E3-1200 v4 CPU with Iris Pro graphics unit is between 35 and 95 Watts. The "Skylake" line of CPUs planned for later this year

with Intel HD Graphics 530 as the first released product derived from an instance of the Intel processor graphics gen9 architecture. According to announcements recently made by Intel such a CPU may come with up to three slices with 24 Iris Pro GPU cores each, providing a total of 72 GPU cores.

	Intel Xeon Processor E3-1200 v3	Intel Xeon Processor E3-1200 v4
Max. System Memory	32GB	32GB
Max. Memory Speed	DD3-1600	DD3-1866
Max. Video RAM	1.7GB	16GB
eDRAM Capacity	N/A	128MB
Graphics Execution Units	20	48

NVIDIA

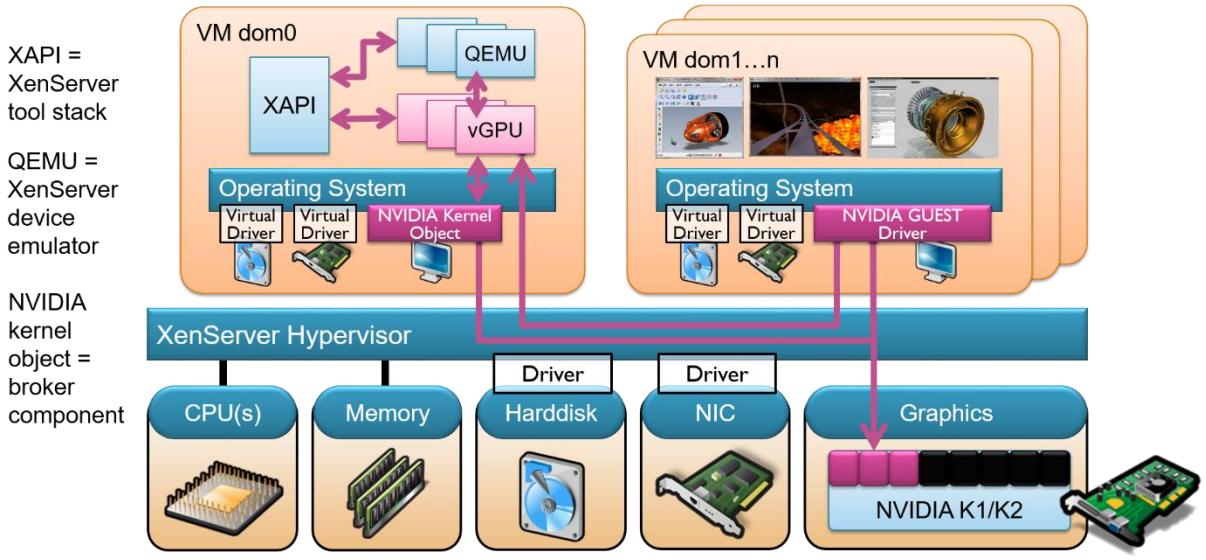
NVIDIA is an US American global technology company based in Santa Clara, California. NVIDIA manufactures graphics processing units (GPUs), but they also have a significant stake in manufacturing system-on-a-chip units (SOCs) for the mobile computing market. In July 2013, NVIDIA joined the gaming industry with its handheld NVIDIA Shield. In addition to GPU manufacturing, NVIDIA provides parallel processing capabilities to researchers and scientists that allow them to efficiently run high-performance applications. They are deployed in supercomputing sites around the world. Besides Advanced Micro Devices (AMD), its other competitors include Intel and Qualcomm.

In September 2013, NVIDIA revealed the Tegra Note 7, a mini tablet computer that runs the Android operating system. This marks a milestone of NVIDIA's move into the mobile computing market, where it produces Tegra mobile processors for smartphones and tablets, as well as vehicle infotainment systems. In addition, NVIDIA released Shield TV and GRID-based virtual gaming service to deliver PC gaming from the cloud.

GRID vGPU

The NVIDIA GRID portfolio of technologies leverages the power of the GPU and graphics applications to deliver GPU-accelerated applications and games over the network to any user. NVIDIA GRID 1.0 GPUs are based on the NVIDIA Kepler GPU architecture. NVIDIA GRID software is a complete stack of GPU virtualization, remoting and session-management libraries that allows multiple users to experience graphics-intensive desktops, applications and games using GPUs. Supported hypervisors are Citrix XenServer and VMware vSphere.

The following image shows the GRID architecture on Citrix XenServer as an example:



NVIDIA GRID technology offers the ability to offload graphics processing from the CPU to the GPU in 3D graphics for virtual desktop environments, allowing the data center manager to deliver true PC graphics-rich experiences to more users for the first time.

NVIDIA GRID 1.0 is available in two models:

- The GRID K1 card is for general purpose VDI sessions. It has four NVIDIA Kepler GPUs (consisting of 768 CUDA cores and 16GB of DDR3 RAM), and NVIDIA figures it's good for 100 VDI sessions on the host
- The GRID K2 card is a special purpose card designed for graphics-intensive environments. It only has 8GB of GDDR5 RAM and two GPUs (instead of four), though they're higher end with 3072 CUDA cores

The GRID vGPU technology is implemented in such a way that it allows multiple resource profiles. The profile defines the number of virtual machines that are assigned to a physical GPU. The amount of video memory is also defined by the profile. Such a GRID-based environment gives administrators a high degree of freedom when assigning graphics resources, allowing them to adapt to changing user requirements.

Today's implementation of the GRID vGPU 1.0 technology supports up to eight users per physical GPU, which equals 32 users on the quad GPU GRID K1 board. Servers typically support two GRID boards, so the maximum is GPU-accelerated 64 users per server. Future versions of GRID vGPU may support higher scalability with just a software update. Of course scalability depends on application usage.

If 3D graphics applications require the maximum amount of GPU memory, then the GRID K1 with 1GB memory and 786 CUDA core is the preferred choice. If a 3D graphics application is consuming more GPU power compared to GPU memory, then the NVIDIA GRID K2 with 8GB memory and 3072 CUDA cores is the preferred choice. Keep in mind when using GRID v1.0 GPU pass-through that the K1 has four GPUs on one board while the K2 has two more powerful GPUs on one board.

NVIDIA GRID 1.0 overview

NVIDIA GRID Graphics Cards	vGPU profile	Graphics memory	Cores per GPU (Shaders)	Maximum number of monitors per user	Maximum resolution per monitor	Number of vGPUs per GPU	Maximum number of users per board	User category
GRID K2 2 phys. GPUs	K280 (PT)	4.096 MB	1542	4	2560x1600	1	2	Power User
	K260Q	2.048 MB	786	4	2560x1600	2	4	Power User
	K240Q	1.024 MB	384	2	2560x1600	4	8	Power User
	K220Q	512 MB	96	2	2560x1600	8	16	Knowledge
	K200	256 MB	96	2	1920x1200	8	16	Knowledge
GRID K1 4 phys. GPUs	Pass-Thru	4.096 MB	192	2	2560x1600	4	4	Power User
	K140Q	1.024 MB	48	2	2560x1600	4	16	Knowledge
	K120Q	512 MB	24	2	2560x1600	8	32	Task Wrk
	K100	256 MB	24	2	1920x1200	8	32	Task Wrk

You can use Process Explorer v15.3 (or later) or GPU-Z to get an approximate understanding of how much GPU core processing power or frame buffer is consumed by an application. This helps you to determine which vGPU profile is best suited for your applications/users or which GRID board to use. Simply run one of these utilities on a physical workstation before you launch a 3D graphics application and take note of the current GPU cycles and GPU system memory utilization before you begin using the application. During the execution of the graphical application you can track the allocation of memory and observe GPU core utilization. This will help you decide on whether the application is more bound by GPU core compute or GPU memory allocation. The performance of one GRID K1 GPU in pass-through mode is comparable with a K600 with more VRAM. The performance of one GRID K2 GPU in pass-through mode is comparable with a K5000.

The Q in vGPU profiles such as K260Q refers to Quadro, indicating that the vGPU profile went through the official Quadro certification process from an ISV perspective. It is key that the selected combination of server hardware and software is officially supported. The combination of server hardware with NVIDIA GRID, the amount of required GRID cards, network card-and storage controller interfaces, Teradici PCoIP server offload card or PCIe flash based storage controllers such as FusionIO need to be supported by hardware vendors such as HP or Cisco. This also includes the virtual desktop software and the selected hypervisor.

GRID 2.0

The new NVIDIA GRID 2.0 introduced at VMworld 2015 still follows the same general GRID philosophy. Like its predecessor, GRID v2 is based on high-end GPU cards, a software layer in the hypervisor and a (certified) graphics driver for the guests. The new GRID solution runs on top of NVIDIA "Maxwell" GPUs that come in two server form factors. The TESLA M6 bare board has one NVIDIA Maxwell GPU, 1,536 CUDA cores and 8 GB of GDDR5 RAM while the dual GPU TESLA M60

card has 4,096 CUDA cores and 16 GB of GDDR5 RAM. GRID v2 enables up to 16 users to share each physical GPU, so like in the previous GRID version the graphics resources of the available GPUs can be assigned to virtual machines in a balanced way.

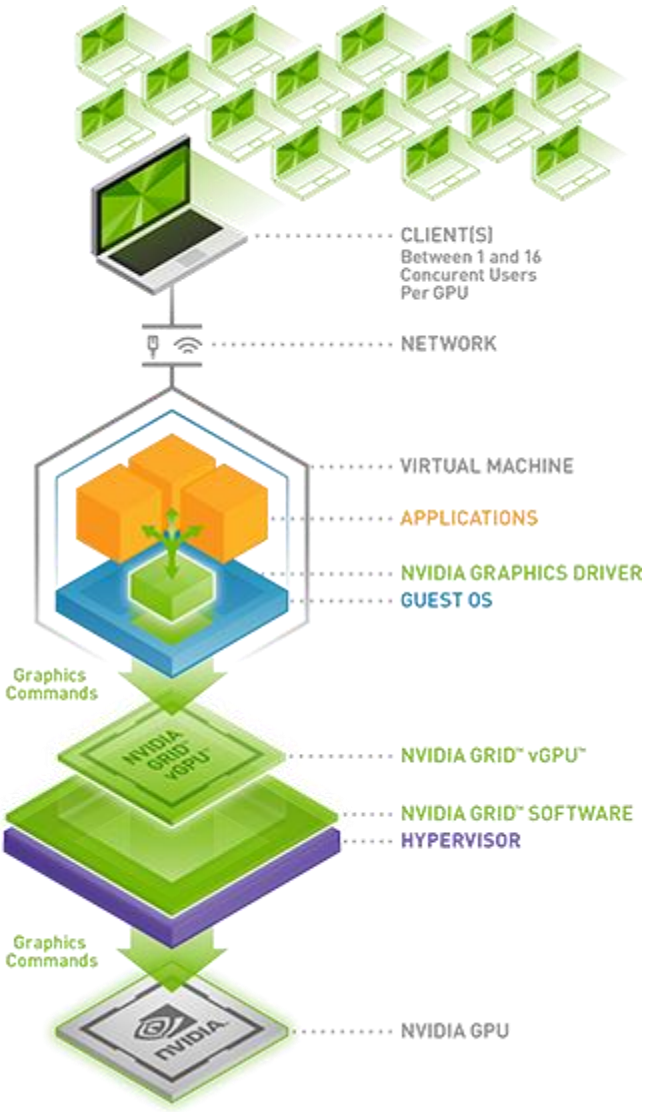


Image source: NVIDIA

What's new is that NVIDIA wants their OEM partners to charge customers a license fee for the new GRID graphics software stack including the GPU broker and the certified driver. The final price per user license will be announced by NVIDIA partners in the near future. The total price for the GRID product is then determined by the number of users or VMs, the selected GRID vGPU profiles and the price for the physical GPU cards. In the table below, you see an overview of NVIDIA GRID K1, K2, M60, M6 and Quadro M6000:

	Tesla M60	Tesla M6	GRID K1	GRID K2	Quadro M6000
GPU	Maxwell	Maxwell	Kepler	Kepler	Maxwell
# GPU	2	1	4	2	1
GPU Type	GM204	GM204	GK107	GK104	GM200
Clock speed	1126 MHz	1126 MHz	850 MHz	745 MHz	988 MHz
CUDA cores – shading units	4096 (2048/GPU)	1536	768 (192/GPU)	3072 (1536/GPU)	3072

	Tesla M60	Tesla M6	GRID K1	GRID K2	Quadro M6000
GFLOPS	4612	4612	3090	3090	6070
Memory size	16GB GDDR5	8GB GDDR5	16GB DDR3	8GB GDDR5	12GB GDDR5
Memory bus	256 bits/GPU	256 bits/GPU	128 bits/GPU	256 bits/GPU	384 bits
H.264 1080p30 streams	36	18	8	12	
Grid vGPU CCU	2/48/16/32	1/2/4/8/16	16	32	N/A
Form Factor	PCIe 3.0 Dual Slot (rack servers)	MXM (Blade servers)	PCIe 3.0 Dual Slot (rack servers)	PCIe 3.0 Dual Slot (rack servers)	PCIe 3.0 Dual Slot (rack servers)
Max. Power	300W	100W	130W	225W	
Cooling	Active/ passive	Bare board	Active/ passive	Active/ passive	Active

NVIDIA GRID 2.0 is available in three editions that deliver accelerated virtual desktops to support the needs of your users. These editions include Virtual PC, Virtual Workstation, and Virtual Workstation Extended. GRID v2 perpetual licenses are sold by Concurrent Connected User (CCU). The following table shows the differences between the GRID 2.0 editions.

	NVIDIA GRID **Virtual PC**	NVIDIA GRID **Virtual Workstation**	NVIDIA GRID **Virtual Workstation Extended**
Maximum number of displays	2	4	4
Maximum resolution per display	2560x1600	2560x1600	3840X2160 (4K)
Windows Guest OS	√	√	√
Linux Guest OS	X	√	√
NVIDIA Quadro software features	X	√	√
CUDA support	X	X	√*
OpenCL support	X	X	√*
GPU pass-through	X	X	√*
vGPU profiles – frame buffer and maximum number of users per GPU			
512MB (16 users/GPU)	√	√	√
1GB (8 users/GPU)	√	√	√
2GB (4 users/GPU)	√	√	√
4GB (2 users/GPU)	X	X	√
8GB (1 user/GPU)	X	X	√

*) CUDA, OpenCL and pass-through only available with 8GB vGPU profile

Infrastructure as a Service (IaaS) with High-End Graphics Support

This chapter focuses on Infrastructure as a Service (IaaS) offering that are available today and provide support for the usage of physical GPUs.

Amazon EC2 - GPU Instances

Within Amazon Web Services (AWS), EC2 (Elastic Cloud Compute) includes GPU-accelerated instance, in particular G2 and CG1. They provide high-end parallel processing capability via NVIDIA GPUs, leveraging scientific, engineering and rendering applications using Compute Unified Device Architecture (CUDA) or the OpenCL parallel computing framework. Other EC2 GPU use cases are the delivery of 3D graphics for virtual desktops, game streaming, 3D application streaming and any other graphics workloads.

The G2 instances provide access to NVIDIA GRID GPUs ("Kepler" GK104) each with 1,536 CUDA cores and 4GB of video memory. The initial NVIDIA driver release provides support for OpenGL 4.3, DirectX 9, 10, and 11, CUDA 5.5, OpenCL 1.1 and GRID SDK. Note: OpenGL doesn't work with the latest G2 drivers from NVIDIA.

A CG1 instance provide access to NVIDIA Tesla M2050 GPUs ("Fermi" GF100), each with 448 CUDA cores and 3GB of video memory. The latest driver release provides support for CUDA 5.5, OpenCL 1.1 and DirectCompute.

The EC2 GPU instances run as Hardware Virtual Machines (HVM-based instances) it uses hardware-assist technology provided by the AWS platform. With HVM virtualization, the guest VM runs as if it were on a native hardware platform, except that it still uses paravirtual (PV) network and storage drivers for improved performance.

These technologies enable Amazon EC2 to provide dedicated access to one or more discrete GPUs in each GPU instance. GPU instances can be clustered and placed into a cluster placement group. Cluster placement groups provide low latency and high-bandwidth connectivity between the instances within a single Availability Zone.

NVIDIA provides Amazon Machine Images (AMIs) for GPU instances based on Amazon Linux and Windows. These reference AMIs include the NVIDIA driver, which enables full functionality and performance of the NVIDIA GPUs. Keep in mind that Amazon EC2 GPU instances may not be available in every region, they must be launched from HVM AMIs, they can't access the GPU unless the NVIDIA drivers are installed and they aren't available for use with Amazon DevPay.

Microsoft Azure - GPU Instances

During the AzureCon online event in September 2015, Microsoft's executive vice president Scott Guthrie announced a new N family of virtual machines with the latest type of NVIDIA GPUs. This was the moment when Microsoft officially entered the market of Infrastructure as a Service for high-end remote visualization. In essence, this means that in the near future Microsoft will be adding NVIDIA Tesla M60 and K80 GPU hardware to their Azure datacenters that will support GPU acceleration for both compute and visualization scenarios.

According to a deep dive session at AzureCon 2015 (https://azure.microsoft.com/en-us/documentation/videos/azurecon-2015-applications-that-scale-using-gpu-compute/) introducing the GPU capabilities on Azure, the core scenarios for this technology are Cloud-based streaming and

gaming, video processing, accelerated desktop applications (OpenGL and DirectX) and GPU compute (CUDA and OpenCL). Even though the NVIDIA M60 GPUs may be similar to those used for NVIDIA GRID, the concept behind the Azure announcement is not based on the NVIDIA GRID software stack. The Azure N-series VMs will also not be based on the para-virtualization concept (API intercept) introduced with RemoteFX v2. Instead, Microsoft decided to implement a GPU pass-through mechanism called Discrete Device Assignment (DDA) for Hyper-V and the Azure core infrastructure and services. In the initial release, Azure N-series VMs will be using the native device driver developed by NVIDIA. This enables full support of DirectX, OpenGL, CUDA and OpenCL.

The following table shows the range of the future N-series VMs with GPUs:

	N1	N2	N10	N11	N12	N21
CPU Cores (E5-2690v3)	6	24	6	12	24	24
RAM (GB	64	256	64	128	256	256
SSD (TB)	~0.5	~2.0	~0.5	~1.0	~2.0	~2.0
Network	Azure	Azure	Azure	Azure	Azure	Azure RDMA
GPU Resources	1 x M60 GPU (1/2 physical card)	4 x M60 GPUs (2 physical cards)	1 x K80 GPU (1/2 physical card)	2 x K80 GPUs (1 physical card)	4 x K80 GPUs (2 physical cards)	4 x K80 GPUs (2 physical cards)

In particular, the N1 and N2 offerings will be best suited for remote visualization, exposing the Tesla M60 GPU. The most interesting use case for such a VM will be running a Personalized Session Desktop with 3D graphics applications, which is basically an RDSH server VM enabled for remote access by only one user. Another use cases is running Citrix XenDesktop with GPU pass-through on N1/N2 VMs.

Enabling Technologies for Remote 3D Graphics

This chapter covers individual technologies that can be used to enable high-end graphics remoting in VDI environments.

HP RGS - Remote Graphics Software

HP Remote Graphics Software (RGS) is the collaboration and remote workstation solutions for serious workstation users and their most demanding applications. All applications run natively on the remote workstation whether it is running Windows or Linux. And RGS works just as well on graphics enabled virtual machines to take full advantage of its graphics resources. RGS works on top of industry standard hypervisors with pass thru graphics as well as virtualized graphics. The desktop of the remote workstation is transmitted over a standard network to a local computer using advanced image compression technology (HP3) specifically designed for digital imagery, text and high frame rate video applications. Redirection of most USB devices to provide an interactive, high performance workstation experience is supported.

Back in 2014 HP introduced HP RGS 7.0 as the first solution delivering workstation-class productivity from Windows tablets by leveraging new touch controls along with HP's industry-leading software. With RGS 7.0, HP turns swipes into hot keys, gestures into control and pinches into zooms, along with a precision onscreen mouse that uses the entire screen as a trackpad. These enhancements increase productivity for remote users of touch PCs or Windows tablets, including the HP ElitePad.

With the new HP RGS version 7.1 HP specifically addresses the market transition to 4k displays. HP RGS 7.1 brings a significant boost in performance, enables 60 frames per second of HD content and delivers smooth 4k remoting of 3D CAD and other graphics heavy applications. In some cases, this is a performance boost of over 2x if compared to RGS 7.0. Also new to RGS 7.1 is remoting of Wacom pen tablets from thin clients to graphics enabled Linux VMs to support digital artists going to a virtual environment.

Designers working for leading automotive, aerospace and architecture companies use HP RGS to circle the globe by taking advantage of the protocol's unique handling of network latency and packet loss, a technology named HP Velocity and AVC feature. Other functionalities of RGS 7 are 10+ display support, fully integrated interactive collaboration, and mathematical and visual lossless compression options. For encoding the data streams, the sender can use an NVIDIA Quadro 2000+ GPU if present, including full OpenGL and DirectX 3D graphics support. Supported connection brokers are Leostream, Dell/Quest and Mechdyne.

The HP Remote Graphics Software architecture is as follows:

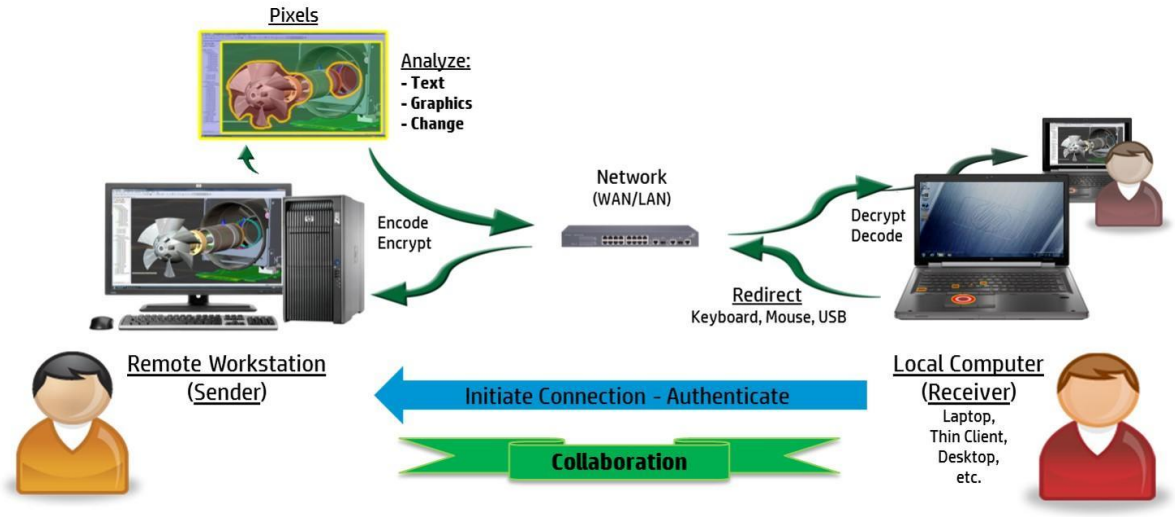


Image source: Hewlett-Packard

One HP RGS 7 license is required for each RGS sender system. HP RGS floating license can be used where licenses are “checked out” on a concurrent use basis up to the number of licenses purchased. In this scenario the use of the license server is required. With HP RGS there are no monthly fees, the Receiver is a free download. As mentioned above, with HP Z workstations the HP RGS sender is included and does not require a separate license.

Teradici

Teradici is a privately held software company and was founded in 2004. It has its headquartered in British Columbia and Santa Clara, California. Between 2004 and 2007, PCoIP (PCoIP) was developed, a remote display protocol available in hardware and software. Its purpose is delivering applications and desktops to PCoIP enabled endpoints. These endpoints can be equipped with a PCoIP interface, such as a Zero client, or they can be software based.

In 2007, the first generation PCoIP processors and workstation products were shipped. In 2008, a strategic licensing and co-development with VMware produces a rich, remote desktop experience. VMware View 4.0 was launched with PCoIP features. In 2011, the Teradici hardware accelerator for VMware View was offered under the name APEX 2800. In 2012, the next generation PCoIP took off named TERA2. In 2013, Amazon selected Teradici PCoIP for Amazon Workspace services, a Desktop as a Service offering.

PCoIP continues to be the foundation for VMware Horizon View and Teradici Remote Workstation solutions. The protocol is based on UDP, uses multiple codecs and is dynamically adaptive, with the rendering happening on the host. Images rendered on the server are captured as pixels, compressed and encoded and then sent to the client for decryption and decompression.

Depending on the nature of the data in the video memory, different codecs are used to encode the pixels. Techniques to compress video images differ in effectiveness compared to those for text. The protocol also dynamically adapts its encoding based on the available bandwidth. In low bandwidth environments it utilizes lossy compression where a highly compressed image is quickly delivered, followed by additional data to refine that image, a process termed "build to perceptually lossless".

The following figure shows a complete overview of the PCoIP hardware and software solutions:

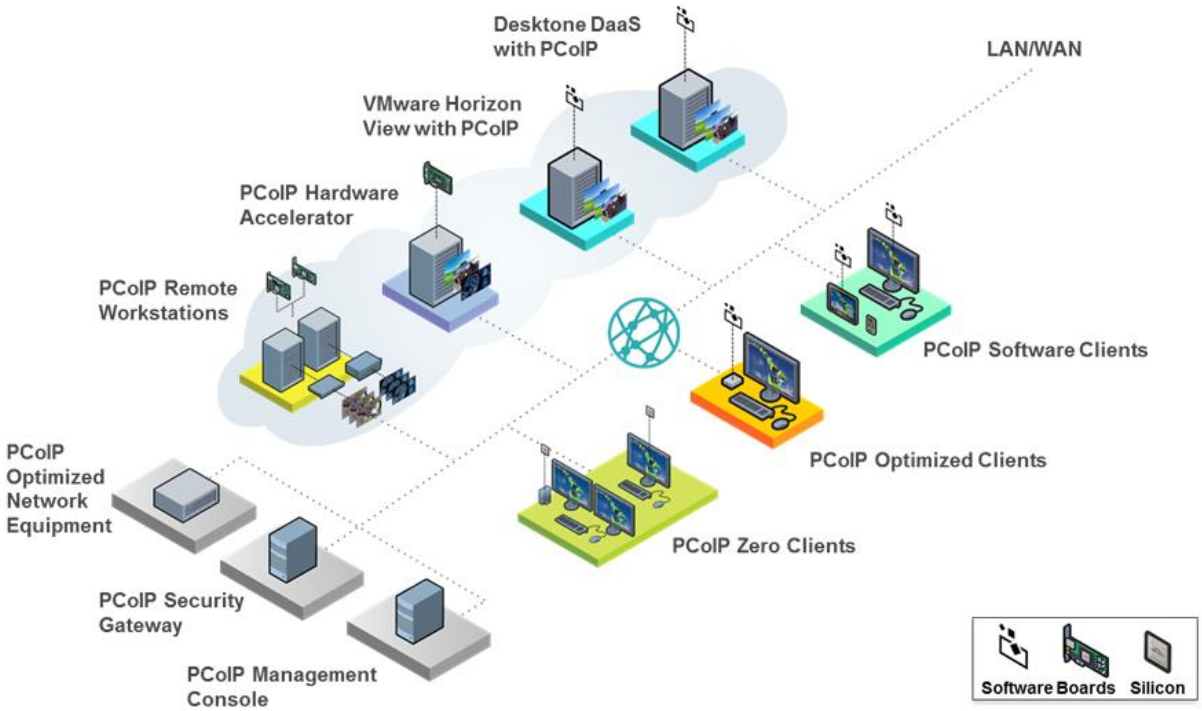


Image source: Teradici

The different high-end graphics scenarios supported by VMware Horizon View and graphics workstations when combined with PCoIP are shown in the chart below.

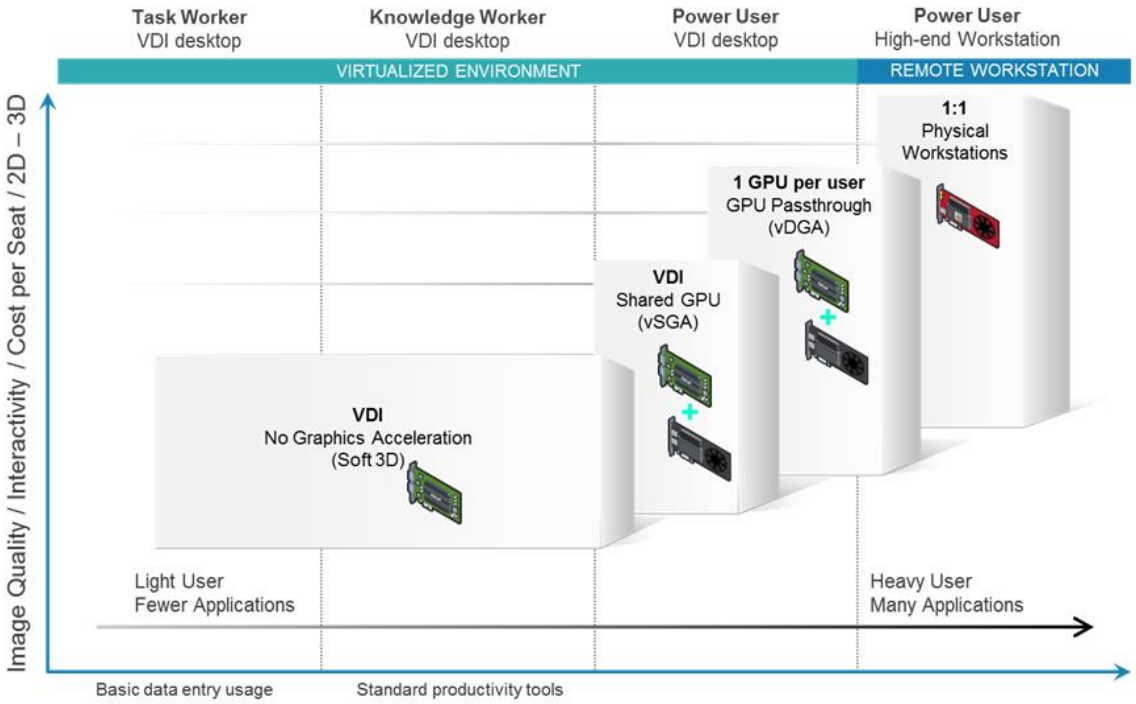


Image source: Teradici

Goliath Monitor

Goliath Technologies' key value is pulling metrics for Citrix XenApp/XenDesktop, End User Experience, NVIDIA GRID vGPU and the supporting IT infrastructure into a single console, to give a broad and deep visibility to preempt potential issues and if they do occur, troubleshoot quickly.

Goliath Technologies is a software vendor headquartered in Philadelphia, USA. Their product Goliath Performance Monitor allows IT professionals to get insights into NVIDIA GRID vGPU and Citrix/VMware environments for proactive performance monitoring and isolation of performance bottlenecks for troubleshooting. This technology gives you real-time visibility into NVIDIA GRID vGPUs graphics card metrics and correlates this data with end user experience, hypervisor performance and XenApp/XenDesktop statistics. The goal is to track performance and analyze utilization trends so you can proactively troubleshoot issues and improve remediation times. This allows you to ensure the proper allocation of vGPU resources to deliver a positive end user experience by preventing overloaded vGPUs and identify sessions impacted by a lack of vGPU, CPU and memory.

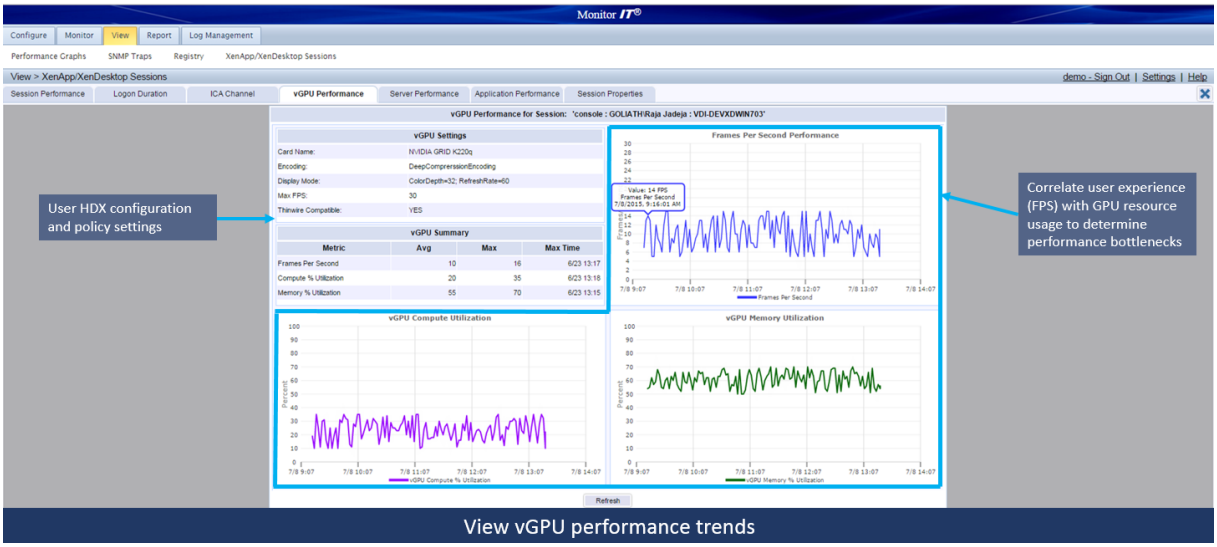


Image source: Goliath Technologies

The key proactive monitoring and troubleshooting features are:

- Real-time NVIDIA GRID vGPU and XenApp/XenDesktop performance metrics
- End user experience monitoring and alerting
- Root cause isolation for troubleshooting
- Real time performance trending dashboards
- Historical reporting and trend analysis

More information: http://goliathtechnologies.com/software/goliath-nvidia-performance-monitor/

Lakeside Software SysTrack

Lakeside Software is a software vendor headquartered in Michigan, USA. Their flagship product SysTrack enables organizations to proactively manage the entire end-user computing infrastructure, including physical PCs, laptops, tablets and VDI environments based on Citrix, Microsoft and VMware virtualization platforms. In particular, it helps customers to leverage hardware accelerated graphics for desktop virtualization by understanding the performance demands of applications in conjunction

with user interaction. By enabling a detailed analysis of an application's graphical demands, SysTrack allows IT professionals to quickly identify which applications benefit the most from hardware accelerated graphics for desktop virtualization.

SysTrack is built on Lakeside Software's patented DataMine technology, which leverages a highly distributed data model. Light-weight software agents are deployed to monitored systems (physical or virtual) and gather up to 10,000 data points every 15 seconds. This is completely transparent to the user and the agent does not require user interaction, reboots, or feature kernel components that could potentially destabilize a system. Once per day, a summary of the collected data is sent to a central master server, where it is stored on a SQL Server backend for further analysis and processing. Alarm states and system state summary data are transmitted to the master server in real time. Overall, the network uses about 100 kB per day per system. Data are collected regardless of whether a system is online or offline and the DataMine technology allows the master to simply catch up when it is next connected. SysTrack includes a collection of analytics tools and visualizers that are sitting on top of DataMine. These tools track which specific data are stored on the central database and which data live on the monitored system.

Once a software product has been fully developed and deployed, it is critical to keep track of the end-user experience to ensure that everything is working as anticipated. This is another area where SysTrack can help, both in capturing the initial end-user experience metrics to benchmark the success of a transformation project and in steady state in the new environment to ensure continued quality. Using proprietary API access provided by NVIDIA, the quality of the end-user experience can now be even more completely characterized with GPU utilization (see figure below). This makes it much easier to pinpoint where oversubscription of resources may be causing issues, and guarantees a basic quality of service to keep productivity at peak levels.

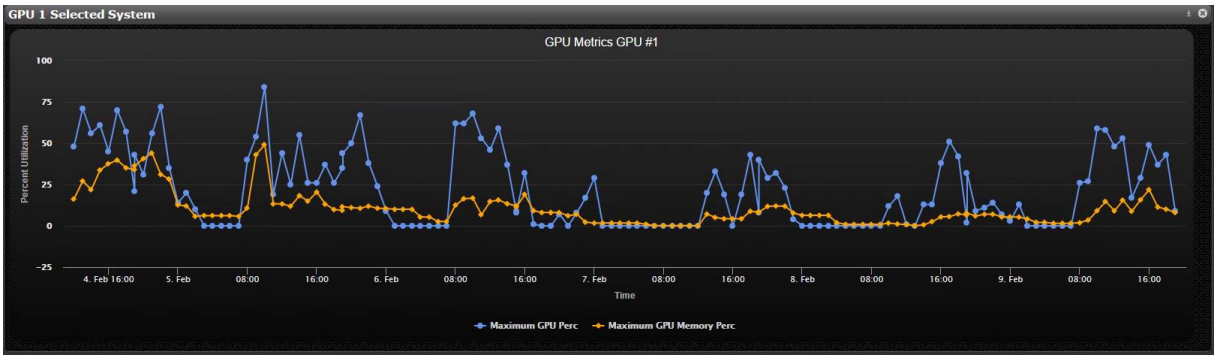


Image source: Lakeside Software

SysTrack continually collects information about software packages as they are used and normalizes all data points for cross platform comparison. In the figure below you can see how the GPU resources consumed by one individual application are tracked. With this information it becomes possible to segment the application portfolio into groupings based on their requirements for specific resources. By tying a general sense of which applications have peak demand to total length of usage it becomes easier to start developing a portfolio made up of different combinations of usage styles. In this context, the grouping of users can be defined by the technical perspective on existing GPU utilization, application portfolio, resiliency and security requirements.

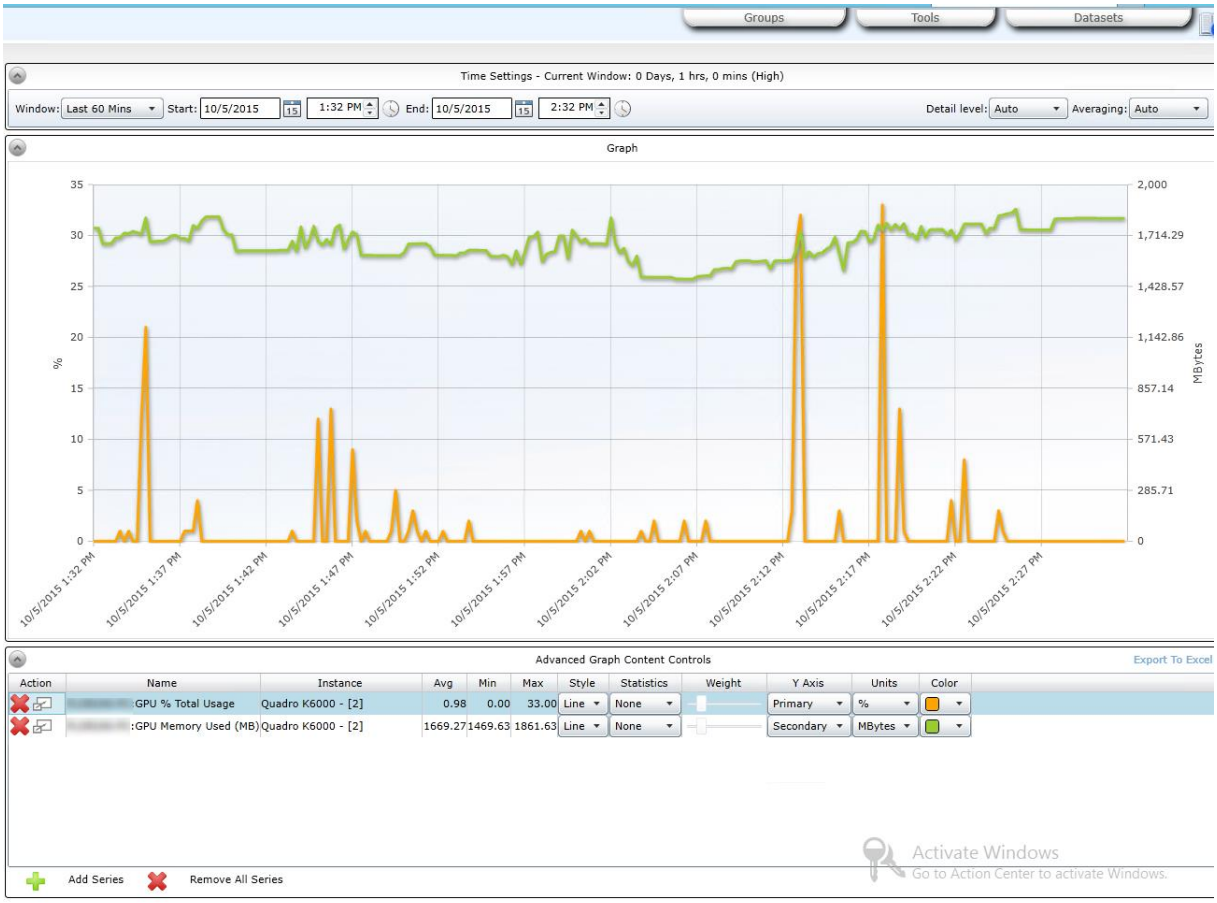


Image source: Lakeside Software

A natural expansion of this is grouping users into distinct workload types to understand how to configure the profile types and GPU assignments for users. Once the target applications and users have been characterized and a plan has been developed it is critical to begin the process of sizing the environment.

By sizing the supporting systems appropriately based on the actual application behavior observed in the environment it becomes much easier to guarantee optimal end-user experience. Through the continual capture of key performance aspects like compute and GPU resource consumption, memory, disk usage, and network utilization a complete, time correlated picture of the expected load in an environment can be created. Different SysTrack Visualizers consume the captured data and provide role-specific analytic views on the quality of the end user experience. Leveraging a data mine that assesses resource demands, application and system reliability and performance, latency to client and backend servers, system events and other criteria, the SysTrack Visualizers measure the extent to which the user's productivity is impacted in terms that relate directly to business cost.

With a complete portfolio plan collected by SysTrack it now becomes possible to move into the next phase and start creating a model for what resources will be required in a complete environment. After each user has been fully characterized throughout the assessment data collection interval, it is possible to use SysTrack's Virtual Machine Planner (VMP) for mathematical analysis, giving insights into infrastructure provisioning.

The NVIDIA Marketplace report created by VMP outlines the number of users that fall into the various use cases (e.g. "high" for a designer or higher end power user), making it much easier to forecast how many users per NVIDIA board can be allocated and in turn how many boards may be

needed. This information creates an easy to use design for a set of user profiles, both for the actual desktop delivery and for the vGPU assignment.

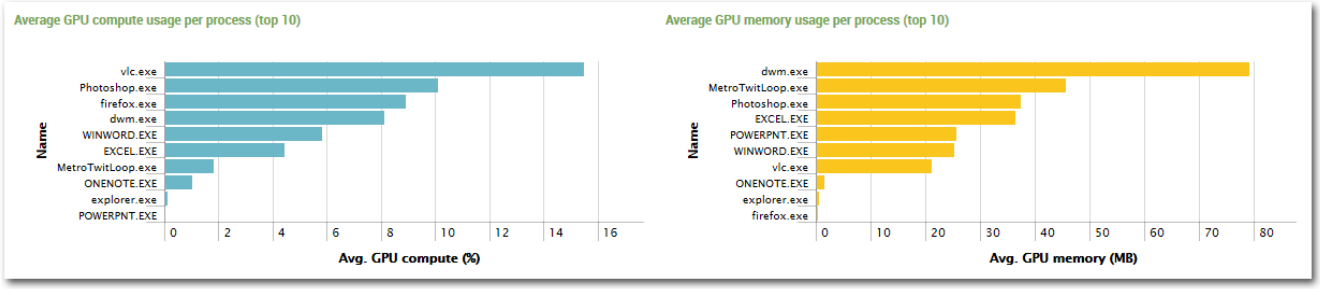
Lakeside Software also provides Marketplace reports for Citrix, Microsoft and VMware VDI infrastructures. But there is an even deeper integration of SysTrack into the most common VDI platforms. SysTrack FastTrack is the virtual desktop assessment tool for Citrix. It helps IT professionals in the transition phase to a user-centric environment by providing the detailed user experience, application and workload analysis to support a successful XenDesktop implementation. SysTrack Desktop Assessment (SDA) for VMware even goes on step further. It is a free virtual desktop and application assessment tool in the cloud, based on SysTrack and hosted on VMware vCloud Air. It allows VMware customers to collect detailed metrics of their own environments and create dynamic reports with deployment and solution recommendations.

uberAgent for Splunk

Splunk is a corporation based in San Francisco, California, with multiple subsidiaries worldwide. The product Splunk Enterprise captures and indexes machine-generated big data in real-time. Data sources are logs, configuration files, messages, scripts and metrics that are produced by applications, servers and network devices. A Splunk Universal Forwarder acts as an agent on the monitored end device and forwards the collected data to Splunk Enterprise for indexing and consolidation. Processing, assessment and analysis of the big data is done with SPL (Search Processing Language). Results are stored a searchable repository from which it can generate graphs, reports, alerts, dashboards and visualizations.

uberAgent (https://uberagent.com/) is a Splunk agent for Windows which implements its own metrics, covering key aspects of user experience and application performance. With the help of this agent important end-user computing metrics can be collected and sent to Splunk Enterprise. uberAgent's high-quality metrics include detailed logon duration and GPU usage, to name just two.

As it becomes more and more common to utilize the GPU for graphics effects, video decoding and even general-purpose computing, administrators can use uberAgent to help them understand how their applications make use of GPU acceleration in order to optimally size the hardware for the workload. GPUs have multiple engines, each serving a different purpose like 2D acceleration, 3D acceleration, video decoding and others. uberAgent determines the utilization of each engine individually. It also reports on shared and dedicated GPU memory usage separately. With that data it is possible to gauge a machine's GPU usage with great accuracy.



uberAgent consists of two components: the endpoint agent collects the data and sends it to Splunk for indexing. The Splunk dashboard app presents the stored data to the user.

The endpoint agent is a lightweight system service. It works on all client and server versions of Windows beginning with Vista / Server 2008, including TS/RDS/XenApp/XenDesktop/Horizon View

and VDI. The dashboard app, on the other hand, contains the dashboards that present the data stored in Splunk. All dashboards are browser-based (no console installation required) and can be modified easily if the need arises.

The entire architecture is extremely flexible and scales from a single server to a globally distributed environment. No database is used; data is stored directly in the file system. Clustering can optionally be enabled for redundancy and high availability. Multi-tenancy is supported.

The primary use cases for uberAgent are resource planning, SLA monitoring, inventory and performance analysis. uberAgent can be licensed per device or per user. Both types of licenses are available as perpetual, one-year term and service provider licenses.

Addendum – 3D Graphics Benchmarks

REUX Tracker

It is not an easy task to measure the quality of perceived user experience in remote session or virtual desktop use cases, and as of today there are no adequate benchmarking methodologies and products available. In order to evaluate the products described in this whitepaper, we had to build our own test lab infrastructure and develop a benchmarking framework called Remote End User Experience (REUX) Tracker.

Our approach uses metrics closely tied to the actual user experience on the client device. Factors such as remoting protocols, host performance, client capabilities, media redirection, network conditions, compression algorithms, media asynchrony and UI response times can impact the user experience, but aren't considered in traditional benchmarking metrics, such as frame rates and system performance counters. In this context, remote end user experience is not only about an application's interface design and usability, it's also about start times, responsiveness of graphical elements and perceived multimedia performance. In addition, as cloud technologies scale out on demand, traditional load tests with growing numbers of user sessions on a relatively static host platform are losing their relevance.

REUX Tracker uses frame grabbers to produce screen recordings, WAN emulators to enforce well-defined network conditions, scriptable remote desktop managers to create user load and monitoring tools to collect correlated sets of telemetry data. Fully automated test sequences simulating users interacting with different applications and media types make sure that all test runs are reproducible. A unique visualization component presents videos and performance data in way that is easy to understand and interpret. SenseConnector and SyncPlayer developed by Sense GmbH are the most important third-party tools integrated into the REUX Tracker methodology.

To visualize and compare the REUX Tracker results we use individual screen recordings as the input material to create pre-defined SyncPlayer configurations in a 4-up split screen arrangement. Four different video sequences play at the same time for visual comparison. The most popular arrangement is called 2x2, where two screen recordings are showing the same sequence of two different test runs in the two upper SyncPlayer quadrants while videos with correlated animated performance counter data are showing in the two lower quadrants.

If you want to learn more about REUX Tracker, SenseConnector or SyncPlayer, send an email to team@teamrge.com.

GPU Tools

- FRAPS: http://www.fraps.com/
- GPU-Z: http://www.techpowerup.com/downloads/SysInfo/GPU-Z/
- GPU Caps Viewer: http://www.ozone3d.net/gpu_caps_viewer/
- GPU Shark: http://www.ozone3d.net/gpushark/
- NVIDIA Inspector: http://download.orbmu2k.de/download.php?id=51
- HWInfo32: http://www.guru3d.com/files_details/hwinfo32_download.html
- GPU Observer Gadget: http://download.orbmu2k.de/download.php?id=49
- GLView (OpenGL Extension Viewer): http://www.realtech-vr.com/glview/download.php
- GLCapsViewer: http://www.saschawillems.de/?page_id=771

- GPUperf2: http://www.virtualexperience.no/sdm_downloads/gpuperf2/
- GPUprofiler: http://www.virtualexperience.no/sdm_downloads/gpuprofiler/
- GPUdetector: http://www.virtualexperience.no/sdm_downloads/gpudetector/
- GPUperfanalyzer: http://www.virtualexperience.no/sdm_downloads/gpuperfanalyzer/
- TessMark for tesselation tests http://www.ozone3d.net/benchmarks/tessmark/
- FluidMark for PhysX tests http://www.ozone3d.net/benchmarks/physx-fluidmark/
- FurMark OpenGL 2.0 Test http://www.ozone3d.net/benchmarks/fur/

Popular Benchmarking Tools

- 3DMark 11: http://www.3dmark.com/de/3dmark11
- Amber GPU Benchmark (CUDA): http://ambermd.org/gpus/benchmarks.htm#Benchmarks
- Maxon CineBench (OGL): http://www.maxon.net/products/cinebench/overview.html
- DaCapo Benchmark Suite, Open Source Java benchmarking: http://www.dacapobench.org/
- FurryBall GPU Benchmark: http://furryball.aaa-studio.eu/products/benchmarks.html
- Futuremark 3DMark (DX): http://www.3dmark.com/), comparison on http://www.futuremark.com
- Futuremark PCMark: http://www.futuremark.com/benchmarks/pcmark
- Geeks 3D MadShaders: http://www.geeks3d.com/madshaders/
- Geeks 3D GpuTest: http://www.geeks3d.com/gputest/
- eDrawings Viewer for SolidWorks: http://www.edrawingsviewer.com/
- Free HDR Images: http://www.hdri-hub.com/hdrishop/freesamples/freehdri
- HDRI Mill: http://www.hdrmill.com/Freebies.htm
- Mars Game (Unreal Engine 3) PhysX benchmark: http://qqsm.zygames.com/v5/benchmark.html
- Shader ToyMark (OpenGL): http://www.ozone3d.net/benchmarks/shadertoymark/
- TessMark for Tesselation Tests: http://www.ozone3d.net/benchmarks/tessmark/
- FluidMark for PhysX tests: http://www.ozone3d.net/benchmarks/physx-fluidmark/
- FurMark OpenGL 2.0 Test: http://www.ozone3d.net/benchmarks/fur/, http://www.majorgeeks.com/files/details/furmark.html
- Luxmark by Luxrender (OpenCL): http://www.luxrender.net/wiki/LuxMark
- Microsoft Process Explorer with GPU monitoring: http://technet.microsoft.com/en-us/sysinternals/bb896653.aspx
- PassMark Performance Test: http://www.passmark.com/products/pt.htm
- Redway3D: http://www.redway3d.com/pages/redsdkDownloads.php
- Redway3D Turbine Demo: http://www.redway3d.com/downloads/public/demos/turbine/turbineDemo.exe
- Redway3D Watch: http://www.redway3d.com/downloads/public/demos/watch/watchDemo.exe
- Revit Forum Benchmark (Logon): http://www.revitforum.org/hardware-infrastructure/1063-rfobenchmark.html
- Unigine Benchmarks (OGL, DX): http://unigine.com/products/benchmarks/
- Unigine Valley: http://unigine.com/products/valley/
- Unigine Heaven: http://unigine.com/products/heaven/
- Unigine Tropics: http://unigine.com/products/tropics/

- Unigine Sanctuary: http://unigine.com/products/sanctuary/
- SPEC ViewPerf 11/12: http://www.spec.org/
  http://www.spec.org/benchmarks.html: 8 important Workstation programs (Catia, Ensight, Lightwave, Maya, Pro Engineer, Solidworks, Siemens Teamcenter, Siemens NX)
  http://www.spec.org/gwpg/gpc.static/vp11info.html
- Siemens TeamCenter Vis:
  https://www.plm.automation.siemens.com/en_us/products/teamcenter/lifecycle-visualization/jt2go/index.shtml
- TeamCenter Models:
  https://www.plm.automation.siemens.com/en_us/products/teamcenter/lifecycle-visualization/jt2go/downloads/index.shtml#lightview-close
- Login VSI & Login VSI Graphics Framework
  http://www.loginvsi.com/documentation/Login_VSI_Graphics_Framework

Free 3D-Applications

- Google SketchUp (http://sketchup.google.com/) - To build models in SketchUp, you draw edges and faces using a few simple tools that you can learn in a short time. With Push/Pull tool you can extrude any flat surface into a 3D form. Furthermore, it works together with Google Earth, that you can import a scaled aerial photograph directly from Google Earth, or use SketchUp to build models which can be seen in Google Earth.
- 3DCrafter (http://www.amabilis.com/products.htm) - 3Dcrafter is a real-time 3D modeling and animation tool that incorporates an intuitive drag-and-drop approach to 3D modeling. The standard version of 3DCrafter is freeware.
- 3Dtin (http://www.3dtin.com/) - The simplest 3D software. You can draw directly from your browser.
- Anim8or (http://www.anim8or.com/main/) - Anim8or is a 3D modeling and character animation program.
- Art of Illusion (http://www.artofillusion.org/) - Art of Illusion is a free, open source 3D modelling and rendering studio. Art of Illusion is more as a 3D design system for animated computer graphics than as an engineering CAD tool.
- Blender (http://www.blender.org/) - Blender is the free open source 3D content creation suite, available for all major operating systems under the GNU General Public License. Blender was developed as an in-house application by the Dutch animation studio NeoGeo and Not a Number Technologies (NaN). It is a powerful program contains features that are characteristic of high-end 3D software.
- BRL-CAD (http://brlcad.org/) - BRL-CAD is a powerful cross-platform open source solid modeling system that includes interactive geometry editing, high-performance ray-tracing for rendering and geometric analysis, image and signal-processing tools, a system performance analysis benchmark suite, libraries for robust geometric representation. BRL-CAD has been the primary tri-service solid modeling CAD system used by the U.S. military to model weapons systems for vulnerability and lethality analyses for more than 20 years. It became an open source project on 21 December 2004.
- Creo Elements/Direct (http://www.ptc.com/products/creo-elements-direct/) - formerly CoCreate - is a complete design environment that offers direct 3D CAD modeler, along with 2D CAD, CAE and integrated product data management (PDM).

- DrawPlus Starter Edition (http://www.serif.com/free-graphic-design-software/?MC=FSSDRAWPLUS) - 100% free and simple, with Accurate vector drawing program, realistic brush, pen, and pencil tools, text on a path, blend modes for advance artistic effects.
- Ensight Free (http://www.ceisoftware.com/ensight-free/)
- FreeCAD (http://sourceforge.net/apps/mediawiki/free-cad/index.php?title=Main_Page) - FreeCAD is a general purpose Open Source 3D CAD/MCAD/CAx/CAE/PLM modeler, aimed directly at mechanical engineering and product design but also fits in architecture or other engineering specialties.
- GIMP (http://www.gimp.org/) – GNU Image Manipulation Program. GIMP is a versatile graphics manipulation package.
- GLC Player (http://www.glc-player.net/download.php) - GLC player is a free application used to view 3d models (COLLADA 3DXML OBJ 3DS STL OFF COFF Format) and to navigate easily in these models. With the album management, capture and multi-capture capabilities, html export and navigation possibilities GLC_Player is the accurate tools to review a lot of 3D models and to create illustrations. GLC_Player is a cross-platform application (Mac, Linux and Windows). It is lighter than regular modelling software so very handy.
- Lego Digital Designer (http://ldd.lego.com/) – 3D program that allows users to build models using virtual LEGO bricks.
- LeoCAD (http://www.leocad.org/trac) - LeoCAD is a CAD program that can be used to create virtual LEGO models. It has an easy to use interface and currently features over 3000 different types of pieces created by the LDaw community.
- Netfabb Studio Basic (http://www.netfabb.com/basic.php) - Netfabb Studio Basic provides mesh edit, repair and analysis capabilities. Its compact size of only a few megabytes allows a quick download, an easy installation and the handling of STL and slice files within seconds.
- K-3D (http://www.k-3d.org/wiki/Main_Page) - K-3D is free-as-in-freedom 3D modeling and animation software. It features a plugin-oriented procedural engine for all of its content, making K-3D a very versatile and powerful package. K-3D excels at polygonal modeling, and includes basic tools for NURBS, patches, curves and animation.
- OpenSCAD (http://www.openscad.org/) - OpenSCAD is a software for creating solid 3D CAD objects. It is free software and available for Linux/UNIX, MS Windows and Mac OS X. it does not focus on the artistic aspects of 3D modelling but instead on the CAD aspects.
- Osamu Mizuno's Metasequoia (http://metaseq.net/en/) – 3D Polygon modeler.
- Tinkercad (http://tinkercad.com/) - Tinkercad is a new and faster way of creating designs for your 3D printer. With only three basic tools you can create a wide range of useful things. Once your project is ready simply download the STL file and start your 3D print.
- Wings 3D (http://www.wings3d.com/) - Wings 3D is a subdivision modeler. It has been developed since 2001. Wings 3D offers a wide range of modeling tools, a customizable interface, support for lights and materials, and a built-in AutoUV mapping facility. There is no support in Wings for animation.
- 3D Slicer is a free, comprehensive software platform for medical image analysis and visualization developed with NIH. supporhttp://www.slicer.org/publications/item/view/2331 and http://download.slicer.org/
- GrabCAD Workbench https://grabcad.com/
- TF3DM http://tf3dm.com/

- 3D Model Free http://www.3dmodelfree.com/
- Free Medical DICOM datasets http://www.osirix-viewer.com/datasets/
- Free HDRI Environments http://www.hdri-hub.com/free-samples & http://www.hdrmill.com/Freebies.htm

WebGL Demo-Websites

- http://alteredqualia.com/three/examples/webgl_materials_skin.html
- http://webglsamples.googlecode.com/hg/aquarium/aquarium.html
- http://middle-earth.thehobbit.com/
- http://www.chromeexperiments.com/webgl/
- http://www.zygotebody.com/
- http://helloracer.com/webgl/
- http://madebyevan.com/webgl-water/