Where Technology Means More®

Delivering Specialized VDI Securely

Produced in partnership with Cisco

Contents

Introduction

A remote workforce strategy has never been more critical to delivering business outcomes. Enabling users that are not only mobile, but long-term remote workers, are table stakes to providing a complete IT package to the business. Companies, like people, come in all shapes and sizes and with very different needs. What remote workforce solutions are leveraged in the overall remote strategic architecture will depend on the needs of the end user.

Virtual Desktop Infrastructure (VDI) has emerged as one of the primary ways companies are enabling their remote workforce, while securing their corporate intellectual property and people. VDI in its basic form has been around since before 2009 and the technology components that support it have gone through their full maturation cycle. Task Worker and Knowledge Worker VDI has prescribed architectures and recommendations that can enable large groups of geographically dispersed remote users.

Problem Statement

Applications running on the client at the end user today expect access to a Graphic Processing Unit (GPU) as a part of the base compute package the endpoint provides. Some basic applications will not install or will only provide limited functionality without a GPU.

For most end users in VDI, the CPU is able to substitute for the GPU without impacting the end user performance. This has cost benefits to the business and is the common VDI solution used for Task and Knowledge Worker users.

Beyond these basic VDI use cases, there is a large section of users that execute applications in their job that require a “real” GPU. The CPU is unable to step in and take the place of the GPU without impacting end user experience. This has caused this group of users to mostly remain physical, leveraging alternate remote work solutions like VPNs.

This deprives a large group of business users from the benefits of VDI and, in turn, prevent the business from leveraging the security and flexibility that VDI provides. In the early days of VDI,

graphically intense applications were a non-starter that slowly became a possibility for those who could afford it or had extremely sharp long-term vision on where the solution space was going. The good news is that the early days are long gone and a mature set of technology tools can now be cost effectively brought to bear and enable this group of users.

It is easy to only see the core components of a VDI architecture. While adding VDI brings benefits across the board, security and geographically dispersed and mobile users add complexity the business needs to address. If Virtual Desktops are designed as non-persistent, meaning they are destroyed at logoff, how does a CISOs team create process to analyze desktops that have been breached? If the virtual machine is deleted at logoff what is left to forensically investigate? Bringing to bear the correct set of tools on this challenge is critical to a complete VDI solution.

For businesses with larger more geographically dispersed users, new challenges are created when users move from geographical area to geographical area. Legacy Global Server Load Balancing (GSLB) is a typical design choice to help with these users, but in a growing software defined world GSLB is beginning to under support the challenges of the mobile user.

Background

GPUs are not new to the VDI architecture space. Major GPU manufacturers have led the way for years, partnering with the leading End User Computing companies to integrate their GPUs into and through their partners hypervisors. These engineering partnerships have provided tools to industry professionals that enable users with GPU needs across the spectrum to be enabled in a cost-effective manner. X86 and Hyperconverged providers, like Cisco, have worked closely with all parties to provide the x86 platform at the core of these functions.

First, a quick level set. In the VDI world there are three high level types of GPU enablement. Direct Pass-Through or Virtual Direct Graphics Acceleration functions as it implies. A GPU is added to the hypervisor host, and then manually passed directly to one virtual machine. In this case a virtual machine running a desktop operating system. This provides the full functionality

of the GPU to the virtual machine. This comes with all the detractors of having physical hardware passed directly to a virtual machine. Most importantly, it was and still is expensive to do. The major end user computing companies do not offer as much automated support for this option.

Virtual Shared Graphics acceleration provides GPUs much the way a hypervisor time slices the CPU. In the hypervisor the same thing is now done for the GPU. The virtual machine is provided a virtual GPU based on the hypervisor manufacturers virtual hardware, and the hypervisor translates calls to the GPU. There is a wide array of GPU support and documentation. See links below. This is a very cost-effective way to provide GPU capability to virtual machines. The important take away is the capability of the hypervisor provided virtual graphics card. Does it support OpenGL 4.1? Does your end user's application require that? Ensuring the application vendors requirements are met by the virtual hardware is table stakes, but not difficult.

For where a virtual software GPU does not support the features needed, or whose ISV mandates a specific GPU vendor supported hardware, there are options. Cisco has partnered with both major GPU manufacturers to ensure support for their proprietary GPU virtualization capabilities. The GPU virtualization methods provided by the two major GPU manufacturers are by far the most functional, but with a cost increase to represent the added capability. Which solution and GPU manufacturer is the correct choice will heavily depend on your application vendor. Cisco ensures that no matter what the application vendor requires, their x86 and hyperconverged platforms can support the required GPUs.

Solution

When identifying users with higher GPU needs, typical current end points providing support to these users are expensive and static to one location. When dealing with large geographically dispersed users' files, it may require replication between sites before the application can reopen the file. Experience shows that these types of applications have their own unique challenges.

The starting point of these solutions is in datacenter architectures. Leveraging a GPU driven VDI solution to enable users in this category removes many of the constraints put in place today by the physical nature of the physical endpoint. Cisco Hyperflex and UCS products offer a flexible range of physical platforms that can support the right number of GPUs per host to properly support this type of architecture.

Now intellectual property can be secured inside the datacenter without replicating and keeping local copies at remote sites. This free's up bandwidth on existing WAN links and allows for enterprise architectures, like next generation Cisco SD-WAN, to properly route virtual desktop traffic from the end users to the datacenter. This is the logical next step beyond legacy GSLB. These SD-WAN links ensure that traffic is properly distributed globally to disperse VDI nodes in the corporate network.

Cisco's SD-WAN architecture directly integrates with Cisco's security portfolio, providing seamless visibility and security for workloads distributed through VDI as the traffic leaves the datacenter. Utilizing Cisco Umbrella, DNS requests are transparently secured while Cisco DUO can be utilized to provide identity services across the enterprise.

Users can access their GPU driven virtual desktops from Apple and Android devices, where before it was not possible. There are many possible benefits, but one unique to medium to high end GPU VDI is possible cost savings. One of the misunderstandings of traditional VDI is that it saves money. VDI is not a cost saving solution, it is a workforce enablement and corporate protection tool. When done correctly, VDI costs approximately the same as standard physical desktops but with all the value of VDI. Physical GPU endpoints that can cost 20-40x what a traditional endpoint costs the company are extremely expensive. When architected correctly, and at the correct scale, GPU enabled VDI can be more cost effective than purchasing and maintaining physical GPU end point workstations.

These types of end point use cases support sensitive intellectual property for the business and directly impact critical business outcomes. To maximize cost efficiency, these virtual desktops are non-persistent to the user. This is the current recommended best practice, but as hinted at

earlier, can pose a problem for the corporate security team to provide forensic investigations after an incident is discovered. Leveraging tools from Cisco like Umbrella and AMP ensure that those forensic tools are available to the CISO without impacting the cost or end user experience of the virtual desktop.

Conclusion

In the end, any business using physical endpoints that are professional grade GPU enabled, may be able to tap into the cost effectiveness of GPU enabled VDI. Regardless of cost, GPU enabled VDI is the most secure way to allow innovation on GPU enabled applications while securing corporate intellectual property and enhancing end user experience and productivity.

GPU driven VDI is leveraged across all sectors of technology from Government to Fortune 100. It is a fully maturated technology solution space that is still fighting its way past the fear, uncertainty and doubt (FUD) of its youth. In a world where remote work is critical, and users on physical GPU workstations are normally critical workers, a GPU enabled VDI solution may no longer be a “nice to have” for a business.

No matter, what it’s worth is checking out.

References

1. NVIDIA GRID - NVIDIA Virtual GPU (vgpu) Software Documentation
2. VSGA - vmware {code}
3. https://www.amd.com/en/graphics/workstation-virtual-graphics
4. https://www.cisco.com/c/en/us/products/hyperconverged-infrastructure/index.html