White Paper | # AMD MULTIUSER GPU: HARDWARE-ENABLED GPU VIRTUALIZATION FOR A TRUE WORKSTATION EXPERIENCE

By Tonny Wong, Product Manager, Radeon™ Technology Group

TABLE OF CONTENTS

Overview

Virtual Desktop Infrastructure (VDI) has evolved over the last few years, enabling richer user experiences and improved manageability and deployment ease. Many traditional VDI enterprise customers have gained productivity and lowered Total Cost of Ownership (TCO) for their desktop users. The growth of VDI needs to address the needs of "greenfield" users, those organizations that want the benefits of secure hosted desktops but with a deployment model that is more consistent with their traditional desk-side workstations. These deployments need to abide to existing datacenter standards for hypervisors while leveraging capabilities that match traditional workstations.

The Trend Toward VDI

Remote graphics protocols have greatly improved user experiences, delivering the feel of a local workstation computing resource for LAN users and optimizing multimedia and graphics capabilities for WAN users. These remote protocols can deliver GPU-rendered content from the datacenter allowing Virtual Machines with standard desktop OS's to be the main deployment method for users of all types. From demanding workstation applications with high 3D GPU needs all the way to standard enterprise desktop users who want GPU-enriched desktop experiences, this range of users can take advantage of a vast array of VDI solutions now in the market.

VDI is a great way to help ensure desktop security by hosting out of an enterprise private cloud (on-premise datacenter) or via offerings from cloud service providers either fully public or via hybrid public/private clouds. However, the capabilities need to match what users expect from their local workstation systems and not limit to a subset of features. Enterprise VDI deployments should have access to GPU resources in the datacenter or service provider that deliver 3D capabilities across many users while still ensuring all graphics API and compute API standards are available, just like on local workstation systems.

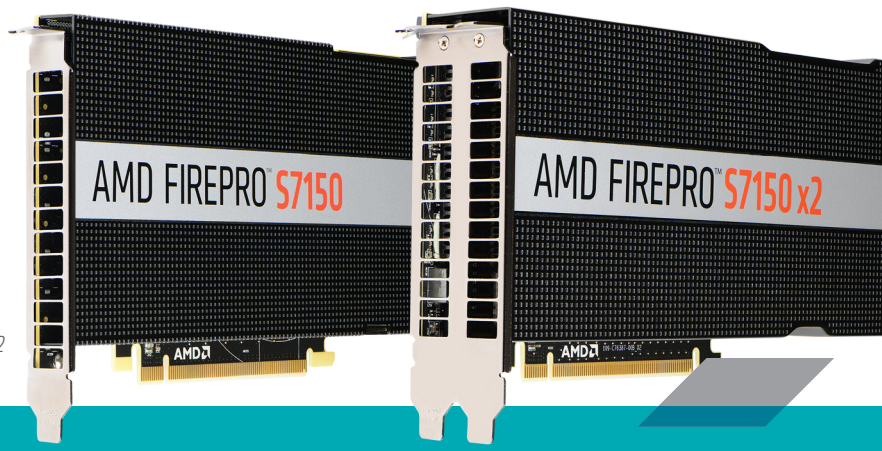
What AMD GPUs Bring to the Virtual Desktop

GPU technology for VDI allows users migrating from physical workstation desktop systems or notebooks to capture the same or better graphics capabilities as their desktop workstation, ensuring productivity while enabling more user types to migrate to VDI. In supporting this migration to VDI, GPU vendors need to ensure that when enabling a GPU for virtualization across many users, this GPU must deliver deterministic performance, helping to better gauge user types and numbers of GPU resources needed.

AMD has spent the last few years implementing features in our GPU hardware to prepare for virtualized platforms. Implementation in our silicon allows our new AMD Multiuser GPU technology to share the GPU resource across multiple users or virtual machines while giving the expanded capabilities users expect from local workstations utilizing discrete GPUs. The AMD Multiuser GPU products can provide enterprise customers with a choice for their GPU and 3D processing needs that can help make GPU use more pervasive on VDI deployments.

VDI with GPUs: Lifting Performance and User Experience

With Virtual Desktop Infrastructure (VDI), one can gain the benefits of security, manageability, remote access to deploy and support enterprise desktop users and may additionally experience lower total cost of ownership (TCO). For the knowledge worker and task worker user types, VDI deployments



The AMD FirePro™ S7150 and S7150 x2

help apply better control of user environments while enabling increased performance by virtue of virtual machines being closer to datacenter, hosted datasets or applications. Users who required higher computing power specifically around GPU technology for 3D and GPU compute applications were either left on physical desktop systems or deployed with comparatively expensive passthrough GPU technology, losing the benefit of distributing the graphics card cost among multiple users. Early virtualized GPU technologies addressed some of these areas by adapting a standard GPU architecture to virtualization via software in the hypervisor, but this isn't the ideal solution to mimic true discrete GPU-like performance. Features like GPU compute functionality are not available, limiting some applications to fallback to CPU usage when a desktop workstation would have leveraged a GPU. Initial pricing for these virtualized GPU solutions was compelling compared to multiple pass-thru GPU devices but they can still have much greater costs than multiple desktop discrete GPUs.

Standard VDI technologies utilize software-emulated GPUs, specifically in VMware vSphere with Horizon View, where the base level graphics capabilities are limited. This works fine for knowledge workers where enabling software 3D emulation with Virtual Shared Graphics Acceleration (vSGA) allows basic applications to run, albeit with higher CPU utilization. vSGA performance is further enhanced by leveraging a hardware GPU with appropriate vSGA drivers from graphics vendors. Even with hardware vSGA support, however, it does not necessarily meet the requirements for more intensive 3D Graphics and Compute user needs. Certifications (CAD/CAE as an example) for applications are not available due to limited support level in graphics APIs like OpenGL® or DirectX®.

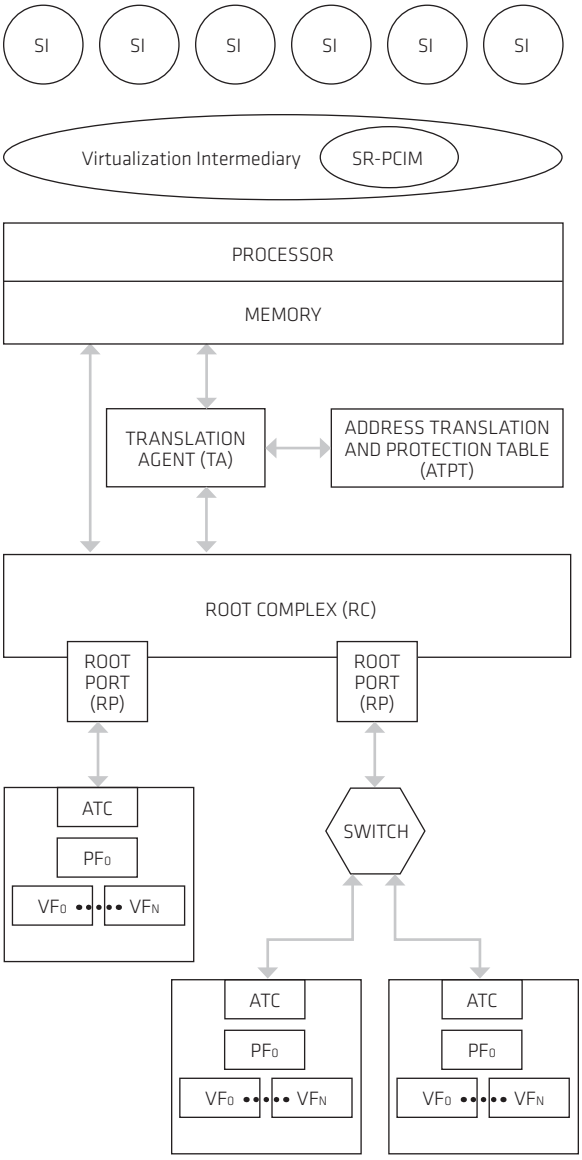
Virtualized GPUs allow workstation and power user categories to migrate to VDI with acceptable GPU performance. Workstation users from CAD/CAE, M&E and specialized segments can leverage workstation-class drivers on applicable platforms to support applications with certification requirements. Power users who rely on DTP/Desktop Publishing, or internal enterprise applications who need GPU support can migrate to VDI environments.

AMD Multiuser GPU – Technology Foundation

Rather than repurposing an existing GPU and adding a software layer to accommodate virtualization requirements, AMD's Multiuser GPU approach is to create an entirely new class of GPU architecture with virtualization capabilities built into the silicon. AMD challenged the notion that the support of GPU virtualization required a proprietary software solution. Compliant with the well-established PCIE® virtualization standard SR-IOV (Single Root I/O Virtualization) specification, AMD has implemented a hardware-based GPU architecture. The culmination of these efforts resulted in the creation of the industry's first hardware virtualized GPU.

The SR-IOV specification defines a virtualized PCIE device to expose one or more physical functions (PF) plus a number of virtual functions (VFs) on the PCIE bus. The specification also defines a standard method to enable the virtual functions by the system software such as the hypervisor or its delegate. These VFs may inherit the same graphics capabilities of the physical GPU, allowing each to become fully capable of supporting the GPU's graphics functionality. Through the PF, system software controls enablement and access permissions of the VFs, internally mapping resources such as the graphics cores and GPU local memory. The task of GPU virtualization management can therefore leverage the existing standard PCIE device management logic in the hypervisor, unburdening the hypervisor from proprietary and complex software implementations. To further simplify the deployment, an optional driver can be loaded to help the hypervisor to enable/disable virtual functions and to manage the Multiuser GPU's resources.

The PF manages sharing of graphical resources by scheduling the GPU cores across VFs and allocating graphics 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 hardware resources and data, at the same time helping keep that data secure. Because each GPU VF is designed to inherit the attributes of the physical GPU, it supports full GPU capabilities allowing the support of graphics and compute features.



SR-IOV Diagram used with permission from PCI-SIG.
Copyright © 2016, PCI-SIG, All Rights Reserved.

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 devices, AMD's native FirePro™ graphics driver that is designed for professional graphics devices can be loaded within the virtual machine to unlock the GPU's graphics and compute capabilities.

A number of FirePro graphics products already support passthrough mode, allowing remote users the ability to access a GPU installed on a host server from a client device. AMD Multiuser GPUs evolved this architecture to support from 1 to 16 VFs, allowing each to appear as a passthrough device with added security and quality of service. Mapping one VF to a virtual machine allows the creation of up to 16 independent guest OSs that are accelerated by a single GPU. User density is limited only by the availability of PCIE slots. Platforms that can support four PCIE slots have the potential of supporting up to 128 users with 128 independent OSs with the dual GPU product.

Key Benefits

Predictable Performance

A key benefit of hardware-based virtualization is that hardware controlled scheduling cycles deliver predictable quality of service (QoS). The fixed scheduling cycles apportioned to each VF ensure that each VF receives its fair share of GPU services.

Predictable performance or deterministic QoS results in smooth transitions from proof-of-concept pilots to organization-wide deployments. Pilot managers determine the capabilities of the GPU during the proof-of-concept phase and scale up or scale down user density (number of users per GPU) as required.

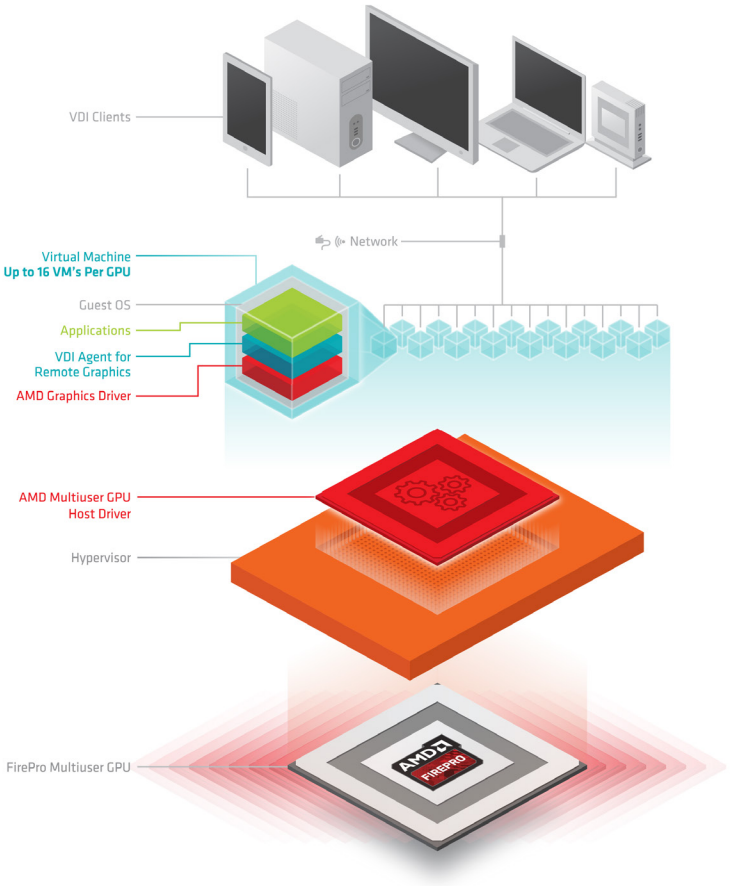
Being able to determine the GPU needs of the user base ties back to an organization's ability to forecast and plan its resources. Under-forecasting results in failing to meet users' performance expectations; over-forecasting results in under-utilizing a configuration.

The predictable nature of AMD's Multiuser GPU solution helps avoid these unwanted outcomes.

Secure Implementation

The push towards virtualization is in part driven by the needs of centralizing and securing data and resources. The cornerstone of AMD's Multiuser GPU technology is its ability to preserve the data integrity of virtualized desktops and their application data. The hardware-enforced memory isolation logic provides strong data security among the VFs, which helps prevent one VM from being able to access another VM's data.

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 own GPU partition, it does not have access to the dedicated local memory of its sibling VFs.



AMD Multiuser GPU Implementation Diagram

Uncompromising Support for APIs and Features

The AMD Multiuser GPU technology exposes all graphics functionality of the GPU to the VF at its partition allowing for not only full support for graphics APIs like DirectX and OpenGL but also GPU compute APIs like OpenCL™. Code written in these standards for the physical device need not be adapted or altered to function in the virtual environment. AMD is the first GPU vendor to support hardware-based native GPU compute features within the virtual environment. Since VFs are allowed access to all of the GPU's rendering resources during their respective time slices, the need to perform post-processing operations to partition data or tasks is not necessary.

AMD operates on the principle of creating customer-centric designs, offering useful features and allowing customers to build usages around these features. Limits are added to control quality, not to constrain utility. FirePro professional graphics, AMD's workstation brand of graphics products, can drive up to six displays per GPU as a standard offering on select AMD FirePro W-series products. Because the Multiuser GPU resides among the FirePro brand of products, the ability to drive up to six displays is an inherent feature. Multiuser GPU products extend this feature by allowing each VF to drive up to six displays within the virtual machine. There is a practical limit when 16 VFs are each given the ability to drive six displays for a maximum of 96 displays. The availability

of local memory will determine the number of displays that will be useful to the end user assigned to the virtual machine. The client device and/or application presenting the virtual machine also has the ability to limit the number of displays supported to control its own quality metric.

Further to the guiding design principle, AMD's Multiuser GPU technology does not require the concept of strict profiles. The host administrator is given complete freedom to configure the GPU for 1 to 16 users and all configurations in between. If the user's graphics rendering needs require an entire GPU then the administrator configures the GPU in direct passthrough mode. If GPU sharing among users is required for cost savings, then the administrator can enable Multiuser GPU to support 2 to 16 users per GPU. A five-user configuration is just as valid as a seven, two or sixteen user configuration.

The Multiuser GPU technology provides the user with close-to-the-metal functionality, bridging the gap to a virtualization experience that is close to being indistinguishable from a native desktop experience.

Simple Approach to GPU Virtualization

Although the Multiuser GPU architecture required years of development, the resulting technology is simple to deploy, making GPU virtualization more accessible to more users. The AMD Multiuser GPU technology is optimized for type 1 hypervisors yet places very little burden on these hypervisors. As long as the hypervisor supports the SR-IOV standard there are no additional requirements from the hypervisor to manage virtualization on this GPU. This simplified approach to unburden the hypervisor from all tasks related to GPU virtualization allows the AMD Multiuser solution to be applied to different hypervisors quite easily. Not only can the technology be applied to different versions of a hypervisor but it can also be ported to other hypervisors with nominal changes required to integrate.

For the user, a simple approach to a complicated technology results in ease of installation and application. If a user is able to configure a device for direct passthrough operation in a one-to-one mapping scenario then it is only two extra steps for this user to load a small driver on the host and configure the Multiuser GPU technology. At that point, all the passthrough GPU virtual devices become available, ready to support the one-to-many (GPU-to-user) mapping usage.

Supported Solutions

AMD continues to invest in the integration of Multiuser GPU technology to leading virtualization applications and hypervisors in the market.

At the launch of the first generation of products with this technology, Multiuser GPU will have support for the VMware product stack. Additional hypervisor support is planned for the product post-launch.

In the guest OS environment, this technology will support 64-bit versions of Windows® 7 and Windows® 8.1 with planned Windows® 10 support post-launch. The driver used in the guest virtual machines is the same standard AMD FirePro graphics driver.

The technology has full support for Horizon View clients (both PCoIP and Blast).Multi displays up to 4 x 4K displays per virtual machines (dependent on client device) are also supported under Horizon View. PCoIP zero clients are also supported as endpoint client devices.

Conclusion

The desire to share storage and network resources sparked innovation of technologies for these devices. The need to centralize all these resources and to secure them in a remote datacenter continues to drive the migration to virtualization. GPU virtualization is a relatively late participant in this migration with early proprietary software-based solutions offering limited GPU capabilities. To become ubiquitous, GPU virtualization technology has to be transparent and standardized, giving users near-desktop experiences without alerting to the fact that they are in a virtualized environment.

AMD Multiuser GPUs push GPU virtualization closer to complete transparency and ubiquity by innovating with a hardware-based solution with conformance to the virtualization industry standard, making it easy to be adopted and integrated into the existing hypervisor ecosystems.

About AMD

AMD (NYSE: AMD) designs and integrates technology that powers millions of intelligent devices, including personal computers, tablets, game consoles and cloud servers that define the new era of surround computing. AMD solutions enable people everywhere to realize the full potential of their favorite devices and applications to push the boundaries of what is possible.

www.amd.com