Bringing New Use Cases and Workloads to the Cloud with Intel® Graphics Virtualization Technology (Intel® GVT-g)

Introduction

From the exponential growth of video on the Internet to desktop virtualization initiatives, media-rich workloads represent a growing share of network traffic. In addition, cloud computing models are incorporating robust media. These usages represent opportunities for businesses to drive new revenue streams and reduce overall costs, but only if the media processing can be managed efficiently. Efficiency in today's world often implies the use of cloud computing, but media workloads have traditionally been difficult to schedule in the cloud due to an inability to access graphics processing unit (GPU) offload capabilities for optimal performance.

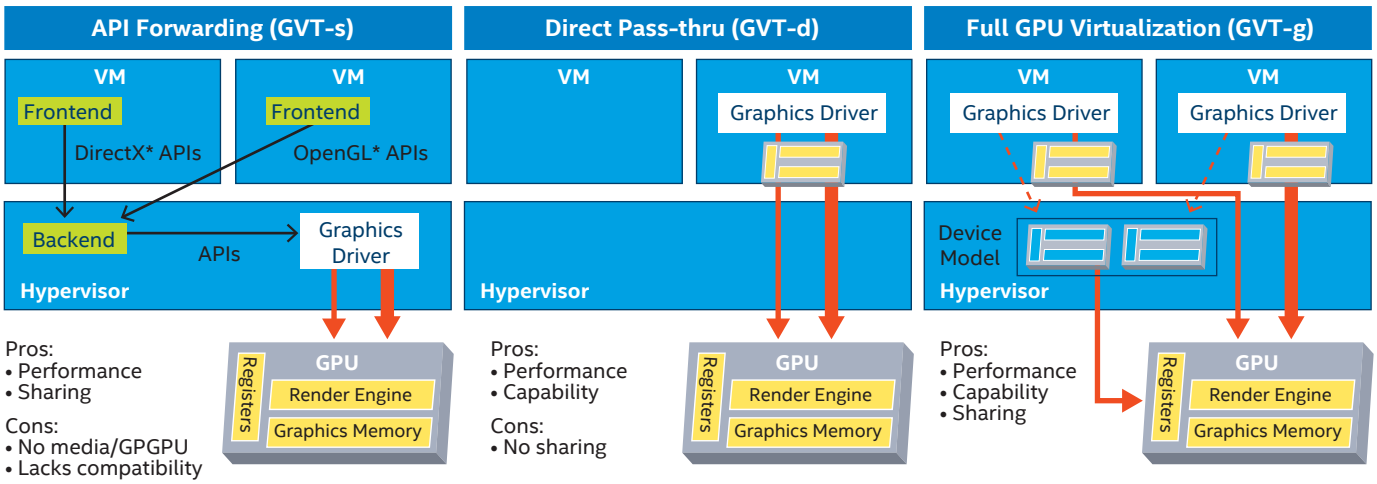
In response to this challenge, graphics virtualization techniques have evolved to allow a media-optimized workload to run on top of a virtualized environment, such as virtualizing the graphics processor for concurrent use to provide direct, dedicated access to a GPU for a single virtualized workload, or providing a pass-through for direct, shared access to the GPU for a number of virtualized workloads.

Choice in Graphics Virtualization Techniques

Intel's comprehensive portfolio of graphics virtualization technologies, known as Intel® Graphics Virtualization Technology (Intel® GVT), encompasses three distinct graphics virtualization approaches, using -d, -s, and -g modifiers to differentiate among them. Developers can select one or more techniques from the Intel® GVT portfolio to best suit their respective solutions and business models.

- Intel® Graphics Virtualization Technology -s (Intel® GVT-s): virtual shared graphics acceleration (multiple virtual machines (VMs) to one physical GPU), also known as Virtual Shared Graphics Adapter (vSGA)
- Intel® Graphics Virtualization Technology -d (Intel® GVT-d): virtual dedicated graphics acceleration (one VM to one physical GPU), also known as Virtual Direct Graphics Adapter (vDGA)
- Intel® Graphics Virtualization Technology -g (Intel® GVT-g): virtual graphics processing unit (vGPU) (multiple VMs to one physical GPU)

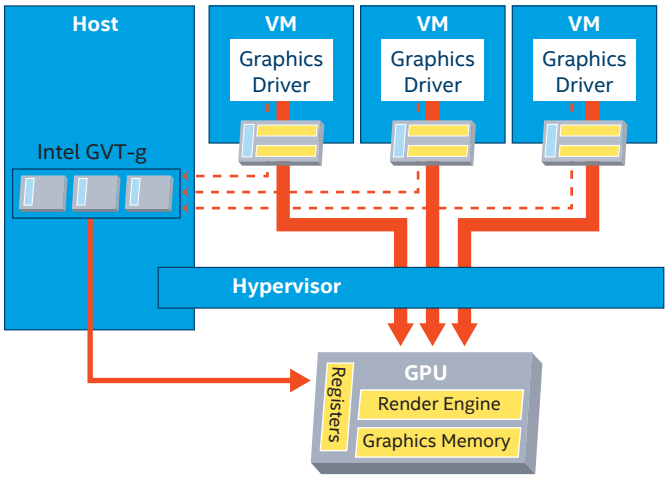
Trade-offs in performance, capabilities, and sharing exist among each of the supported types of graphics virtualization, where performance refers to direct GPU acceleration, capability indicates consistency in visual experience, and sharing denotes multiple VMs. The API forwarding approach offered by Intel GVT-s is capable of supporting an arbitrary numbers of guests via APIs, but this approach limits the versatility of the GPU in cases where support for media or general-purpose GPU workloads is needed. In terms of raw performance, Intel GVT-d, with direct pass-through to the GPU, offers the lowest overhead and may be ideally suited for environments with small numbers of GPU-intensive workloads or with individual workloads capable of fully consuming all available GPU resources. The full GPU virtualization provided by Intel GVT-g offers a solid balance of performance, capabilities, and sharing, allowing near-native performance and full binary compatibility for up to seven guests per server.



Harnessing Near-Native Experiences with Full GPU Virtualization

Intel GVT-g offers a full GPU virtualization approach with mediated pass-through support for Intel® Processor Graphics. This approach maintains a virtual GPU instance for each VM, with direct assignment of performance-critical resources. Running a native graphics driver inside a VM (without hypervisor intervention in performance-critical paths) delivers an optimum balance of near-native performance, capabilities, and sharing. Armed with the ability to take optimum advantage of Intel Processor Graphics within VMs, cloud service providers can deliver cloud solutions optimized for visual computing that offer experiences equivalent to running on bare metal. This approach is supported starting with 4th generation Intel® Core™ processors, with a solid GPU virtualization roadmap on the horizon.

In addition, in the spirit of its upstream-first philosophy as a member of the open-source community, Intel works upstream to ensure that full, open-source implementations of Intel GVT-g technology exist for open-source virtualization hypervisors, Kernel-based Virtual Machine (KVM) and Xen*, known as KVMGT and XenGT respectively. KVMGT and XenGT deliver excellent virtual GPU performance into VMs, delivering > 85 percent 3D performance in VMs compared to native.[1] These implementations support up to seven virtual GPUs in parallel on 5th and 6th generation Intel® Core™ processors (Intel® Core™ i3, Intel® Core™ i5, and Intel® Core™ i7 processors) with Intel Processor Graphics as well as Intel® Xeon® processors E3 v4 family with Intel Processor Graphics.



Benefits

- Provides near-native performance for visual computing tasks inside a VM by allowing direct hardware execution of performance-critical operations
- Delivers Intel Processor Graphics hardware acceleration benefits to multiple VMs simultaneously, with zero porting effort needed by the customer to migrate visual computing tasks to the cloud
- Supports up to seven virtual GPUs in parallel on 5th and 6th generation Intel® Core™ processors with Intel Processor Graphics as well as Intel® Xeon® processors E3 family with Intel Processor Graphics, with more VMs possible in future generations
- Offers total cost of ownership (TCO) savings on Intel Xeon processor E3 family-based platforms with Intel Processor Graphics as well as the full, open-source implementation of Intel GVT-g technology

Usages

- Media decode/encode/transcode in the cloud, enabling greater flexibility and improved resource utilization in balancing media workloads without relying on expensive, single-purpose DSP equipment
- Visual understanding workloads, in which traditional computer vision techniques are augmented with big data analytics and machine learning
- Virtual Desktop Infrastructure (VDI) that can support remote desktop sharing from thin clients, while running heavy GPU workloads on a server, such as remote desktops and workstations for high-demanding workloads, graphics editing, drafting, and CAD software
- In-vehicle infotainment (IVI), combining the dashboard and entertainment systems onto a single CPU/GPU to reduce cost and complexity
- Classroom e-learning, with benefits similar to VDI for license consolidation and optimized resource utilization

To learn more about Intel® Graphics Virtualization Technology (Intel® GVT), visit:

https://01.org/igvt-g
http://www.intel.com/visualcloud

[1] Server: Intel BDW-H SDP. Processor type: Intel® Core i7-585HQ @ 2.70 GHz. GPU Model: Iris Pro 6200 GT3 (0x20). Memory: 16GB 1600 MHz. BIOS settings: HT Enabled: Yes. GTT Size: 8MB. Aperture Size: 1024MB. RC6: Enable. C State: Enable.
Host Operating system/Kernel: Ubuntu 14.04.3/drm-intel 4.3.0-rc6. Guest Operating system/Kernel: Ubuntu 14.04.3/drm-intel 4.3.0-rc6. Software and workloads used in performance tests may have been optimized for performance only on Intel® microprocessors. Performance tests, such as SYSmark* and MobileMark*, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. Results have been estimated or simulated using internal Intel analysis or architecture simulation or modeling, and provided to you for informational purposes. Any differences in your system hardware, software or configuration may affect your actual performance.

Printed in USA 0516/NKR/MESH/PDF ♻ Please Recycle