GPU VIRTUALIZATION BREAKS NEW GROUND IN TRAINING & SIMULATION

GPU virtualization is gaining the attention of OEMs in training and simulation. It's a new but proven approach, and stands to transform deployments by reducing system footprint and simplifying long-term system management.

By Jeff Krueger

Training and Simulation – Market Snapshot

Housing massive deployments of simulators and classroom trainers, simulation facilities face power and cooling demands similar to a data center. Coupled with complexity of updates for individual systems powered by individual servers, operating costs and maintenance resources are significant and steadily increasing. At the same time, solutions must keep pace with the quality and performance of graphics advances such as augmented and virtual reality (AR/VR) while remaining interoperable, global, and highly secure. These challenges are sparking the conversation about GPU virtualization. By enabling OEMs to meet long-term goals of reducing footprint and operational costs, GPU virtualization represents transformation for the industry.

Introducing GPU virtualization in Training and Simulation

While hardware acceleration through CPU virtualization is a long-solved challenge for OEMs, GPU virtualization is breaking new ground. In the training and simulation market, this evolving technology is a possible fit for OEM applications. Developments in GPU performance are not only advancing GPU virtualization as a viable strategy, but also highlighting its capacity for reducing system footprints and creating long-term advantages in overall system management. As the market evolves with training scenarios that are continually more immersive and sophisticated, OEMs must embrace innovations that appeal directly to these key market drivers. So, what are the options for GPU virtualization? And how can OEMs avoid common design pitfalls? This paper provides insight that can help OEMs find new competitive value, embracing virtualized GPU technologies to drive next generation training and simulation systems.

Understanding the Possibilities

Training and simulation developers face a constant challenge in balancing system performance with footprint and serviceability. Systems often have higher-level security requirements, with unique demands that add complexity and cost. PCoIP has answered the need, as a non-virtualized and 1:1 model where each rendering node is securely connected to its own corresponding rack system. Yet PCoIP is hardware-heavy and costly. Each thin client, its PCoIP card, and networking requirements expand the challenge in implementing and managing more systems.

GPU virtualization changes this landscape dramatically. Today, multiple image generators (IGs) or render nodes within a simulator can be condensed into a single physical system. Instead of numerous systems performing individually, numerous virtual machines (VMs) are fed by just one larger system. Techniques vary based on the application's requirements, and include CPU/GPU passthrough and vGPU (virtual graphics processing unit) strategies. Both consistently capitalize on GPU advancements to optimize performance, reduce footprint, and simplify system management. Depending on how the IG is configured, multiple rendering nodes can be placed in a single 4U chassis, reducing overall footprint of the IG.

Virtualizing with Passthrough

Passthrough virtualization is about more than just virtualized GPUs, even though most publicly available documentation calls out only GPU-related design options. In practice, any PCIe (Peripheral Component InnerConnect Express) component or USB device can be passed through and its resources divided in a VM instance. This allows all the hardware resources, including the GPU and other hardware-based resources, to be fully utilized by individual VMs; each VM can use some, all, or none of the cards available for passthrough. This enables high-speed I/O, as fast as physical systems, while hardware demands are kept lightweight with a significantly reduced footprint.

To prove the concept, two virtual machines were set up on the same host, each having access to import a USB 3.0 I/O card and an NVIDIA Quadro M6000. Using 3DMark Time Spy testing, GPU benchmarks demonstrated minimal performance impact as a VM, even while

the USB 3.0 card was used to drive complicated I/O such as an Oculus Rift. Click here for the entire benchmark document.

Choosing vGPU

vGPU is an alternative to the passthrough protocol for virtualization. It works by creating a profile on the host GPU and operates in concert with CPU virtualization. In this design, the total performance of the host GPU can be shared by VMs using their own resource pool. vGPU strategies are typically used to condense systems with low- to mid-level graphics and latency requirements into a very dense footprint. These applications can be illustrated by out-of-window graphics – such as helicopter window views, knee or dashboard views, or landing gear cameras – which require less GPU performance than higher resolution views that are in-frame. Ignoring CPU constraints, there could be 100 or more VMs designed into a single box, creating a system with very few failure points to manage. Storage and processing can easily be moved from inside a classroom to a secure server room, meeting a critical need for military training environments.

vGPU performance was benchmarked using Unigine Heaven, a benchmark tool without significant dependence on CPU or memory performance. The initial evaluation ran eight VMs on a single host with two NVIDIA® Tesla® M60 cards in three different scenarios. Note the Tesla M60 is based on the same silicon as the NVIDIA GTX 980. With two profiles running, each GTX 980 performed equivalent to a GTX 950; with three profiles running, the GTX 980 performed equivalent to a GTX 750. Overall, results indicated up to 16 clients can be placed in a single chassis with no impact on graphics performance beyond the impact of running two clients on a single GPU. Click here for the entire benchmark document.

Capitalizing on Centralized Storage

Both passthrough and vGPU strategies also enable access to a single centralized storage solution, maintaining all images in one repository. Dedicated Computing's ZetaFlex illustrates this type of attached storage appliance, which is optimized for faster application performance through low latency and high bandwidth. ZetaFlex's proprietary software is customizable for read-heavy or read/write balanced workloads. Users can capitalize on storage-compute proximity that eliminates the unpredictable nature of remote network connections.

Tapping into the Art of GPU Virtualization

Virtualizing GPUs is a challenge, particularly in graphics-intensive applications such as training and simulation. Without specific expertise, it's common to vastly over- or under-estimate what it can do for your application. For example, the flexibility of creating a greater number of clients requires expertise in provisioning CPU resources to optimize system I/O.

It's also possible your training and simulation application might be most effective using both vGPU and passthrough models. Consider a high-fidelity cockpit simulator offering instrumentation panels. With comparatively low-render requirements, instrument panels are ideal for vGPU virtualization, while

higher-level graphics of the main render IGs are handled by passthrough. This is entirely customer-dependent and demands a deeper level of virtualization expertise.

Depending on your application, passthrough or vGPU techniques may provide the greatest value, and both enable connection to local, high performance storage for a single image repository.

Partnering for Success with Dedicated Computing

As an Original Design Manufacturer (ODM) of proprietary, highly-engineered computing systems, Dedicated Computing is a partner to OEMs in training and simulation markets.

We're problem-solvers by nature, and our team recognizes that virtualization is more complex than delivering a specific type of hardware, amount of RAM, or compute or graphics processor. Establishing a target frame rate or performance goal are strategies we're comfortable discussing. We're here to bring a fresh perspective to your lifecycle management, field management of deployed systems, and new paths to revenue.

Given the swift evolution toward new forms of content such as serious gaming and AR/VR, Dedicated Computing understands that OEMs need creative solutions that reduce footprint and simplify system management, even as they ensure performance and uptime.

To connect with the Dedicated Computing team for support in design, development, or deployment, call 877.333.4848 or connect via email at support@dedicatedcomputing.com.

About the Author

Jeff Krueger is Dedicated Computing's Director of Systems Engineering. He joined the company in 2011. Krueger is responsible for the strategic direction of researching and developing new products and technologies that meet market demand and keep the company competitive. He brings 29 years of hardware and software engineering and product development experience.

His career history includes various roles within engineering, R&D, product management, project management, and data center operations. Krueger holds a BS degree in Computer Engineering from the Milwaukee School of Engineering, a MBA from the University of Wisconsin - Oshkosh and a MS in eBusiness from the University of Wisconsin - Milwaukee. Additionally, he is a member of the Software Engineering Industry Advisory Board for the Milwaukee School of Engineering and is an active mentor and sponsor of FIRST Robotics.