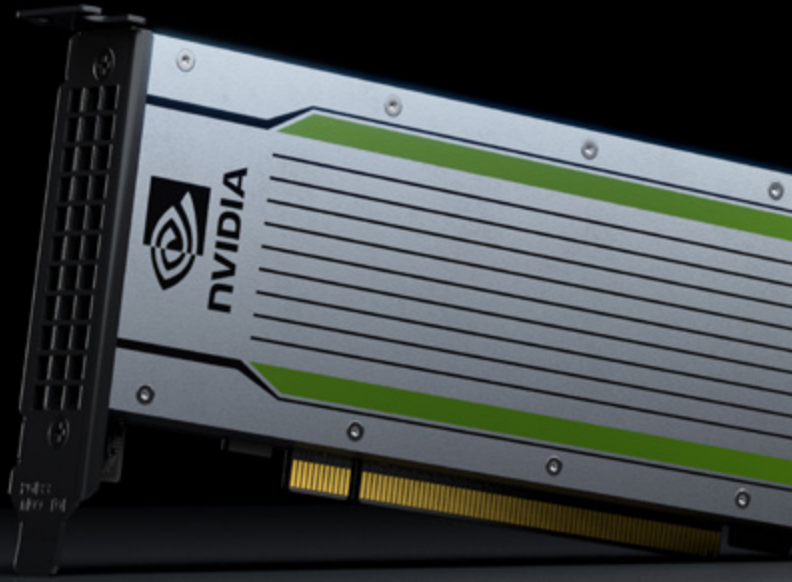# NVIDIA T4
FOR VIRTUALIZATION

Flexibly Power Any Virtual Workload

The NVIDIA® T4 data center GPU is well suited to run a variety of virtualized workloads. Based on the latest NVIDIA Turing™ architecture, the NVIDIA T4 Tensor Core GPU is one of the most flexible GPUs to date, capable of running any workload. By providing the computation required to deliver real-time ray tracing, the same GPU computing platform used by designers and engineers can now also be used by artists to create photorealistic imagery that features light bouncing off surfaces just as it would in real life. This computing platform also enables AI-enhanced graphics, video, and image processing to speed creative workflows, as well as Tensor Cores to accelerate deep learning inferencing workloads. And as the modern digital workplace becomes more graphics intensive, this platform can be used to provide a native-PC user experience for virtualized office productivity applications. With support for Live Migration of GPU accelerated VMs and other high performance technologies, this universal GPU accelerator provides the ultimate flexibility for cost-effectively scaling VDI.

SPECIFICATIONS

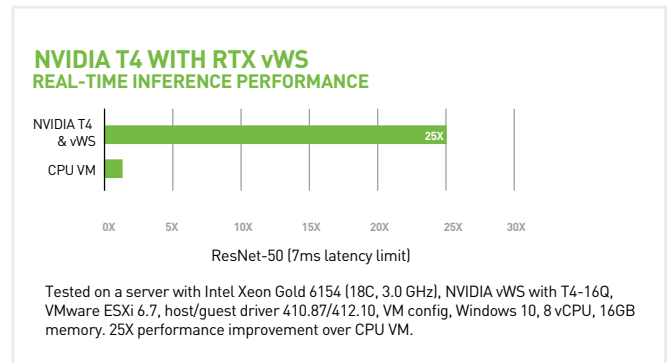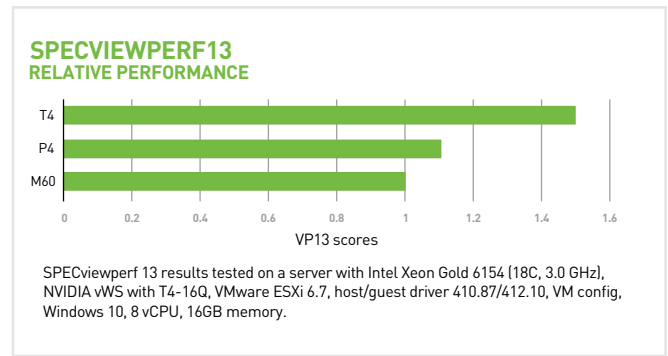|	
GPU Architecture	**NVIDIA Turing**
NVIDIA Turing Tensor Cores	**320**
CUDA® Cores	**2,560**
RT Cores	**40**
Memory Size	**16GB GDDR6**
Memory BW	**Up to 320GB/sec**
vGPU™ Profiles	**1GB, 2GB, 4GB, 8GB, 16GB**
Form Factor	**PCIe 3.0 single slot (half height & length)**
Power	**70W**
Thermal	**Passove**

Faster Virtual Workstation Performance

With the T4 and NVIDIA RTX™ Virtual Workstation (RTX vWS) software, virtual workstation users can achieve up to 2X performance (compared to the M60) and access a 16GB framebuffer (double the P4), which means they can work on larger models and achieve their best designs, faster.

T4 powered RTX virtual workstations running deep learning inferencing workloads can perform up to 25X faster than a VM driven by a CPU-only server. NVIDIA GPU Cloud™ (NGC) containers simplify the installation process for IT and reduce the risks of implementing deep learning workloads.

The T4 is an RTX-capable GPU, supporting the enhancements of the RTX platform. When combined with RTX vWS, virtual workstations can achieve real-time ray tracing performance. Bring creations to market faster with accelerated batch rendering, and speed creative workflows with AI-enhanced denoising. With RTX, artists working in RTX virtual workstations can create photorealistic designs with accurate shadows, reflections, and refractions and can do so on any device, from anywhere.

Power Efficient Virtual Desktops

Knowledge workers using modern productivity applications and viewing video on multiple and 4K

SPECVIEWPERF13
RELATIVE PERFORMANCE

SPECviewperf 13 results tested on a server with Intel Xeon Gold 6154 (18C, 3.0 GHz), NVIDIA vWS with T4-16Q, VMware ESXi 6.7, host/guest driver 410.87/412.10, VM config, Windows 10, 8 vCPU, 16GB memory.

NVIDIA T4 WITH RTX vWS
REAL-TIME INFERENCE PERFORMANCE

Tested on a server with Intel Xeon Gold 6154 (18C, 3.0 GHz), NVIDIA vWS with T4-16Q, VMware ESXi 6.7, host/guest driver 410.87/412.10, VM config, Windows 10, 8 vCPU, 16GB memory. 25X performance improvement over CPU VM.

monitors require increased graphics support to enjoy a native-PC user experience in a virtual environment. With its 16GB framebuffer and compact form factor, two T4 GPUs can deliver the same user density as a single M10 with a 32GB framebuffer. Because the T4 consumes only 70W of power, it can deliver this density with lower power consumption. The T4 also supports VP9 video decode and H.265 (HVEC) encode/decode and provides more than 1TB of system memory. For data centers tasked with running a wide range of graphics and compute intensive workloads, the flexibility of T4 makes it an ideal solution.

To learn more about NVIDIA virtual GPU technology, visit **www.nvidia.com/virtualgpu**