TESLA T4 FOR VIRTUALIZATION

NEW GENERATION OF COMPUTER GRAPHICS DRIVES INCREASED VERSATILITY AND UTILIZATION

Technology Brief | January 2019

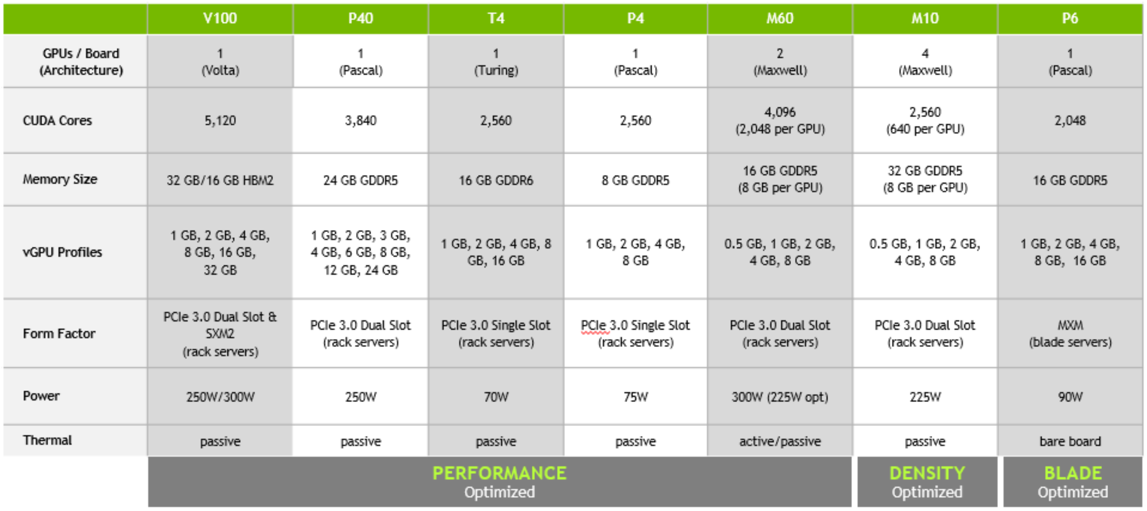
By Emily Apsey, NVIDIA Technical Marketing Engineer

POWERING ANY VIRTUAL WORKLOAD

The NVIDIA® Tesla® T4 GPU, based on the latest NVIDIA Turing™ architecture, is now supported for virtualized workloads with NVIDIA virtual GPU (vGPU) software. Using the same NVIDIA graphics drivers that are deployed on non-virtualized systems, NVIDIA vGPU software provides Virtual Machines (VMs) with the same breakthrough performance and versatility that the T4 offers to a physical environment.

NVIDIA initially launched T4 at GTC Japan in the Fall of 2018 as an AI inferencing platform for bare metal servers. When T4 was initially released, it was specifically designed to meet the needs of public and private cloud environments as their scalability requirements continue to grow. Since then there has been rapid adoption and it was recently released on the Google Cloud Platform. The Tesla T4 is the most universal GPU to date -- capable of running any workload to drive greater data center efficiency. In a bare metal environment, T4 accelerates diverse workloads including deep learning training and inferencing. Adding support for virtual desktops with NVIDIA GRID® Virtual PC (GRID vPC) and NVIDIA Quadro® Virtual Data Center Workstation (Quadro vDWS) software is the next level of workflow acceleration.

The T4 has a low-profile, single slot form factor, roughly the size of a cell phone, and draws a maximum of 70W power, so it requires no supplemental power connector. This highly efficient design allows NVIDIA vGPU customers to reduce their operating costs considerably and offers the flexibility to scale their vGPU deployment by installing additional GPUs in a server, because two T4 GPUs can fit into the same space as a single Tesla M10 or M60 GPU, which could consume more than 3X the power.



	V100	P40	T4	P4	M60	M10	P6
GPUs / Board (Architecture)	1 (Volta)	1 (Pascal)	1 (Turing)	1 (Pascal)	2 (Maxwell)	4 (Maxwell)	1 (Pascal)
CUDA Cores	5,120	3,840	2,560	2,560	4,096 (2,048 per GPU)	2,560 (640 per GPU)	2,048
Memory Size	32 GB/16 GB HBM2	24 GB GDDR5	16 GB GDDR6	8 GB GDDR5	16 GB GDDR5 (8 GB per GPU)	32 GB GDDR5 (8 GB per GPU)	16 GB GDDR5
vGPU Profiles	1 GB, 2 GB, 4 GB, 8 GB, 16 GB, 32 GB	1 GB, 2 GB, 3 GB, 4 GB, 6 GB, 8 GB, 12 GB, 24 GB	1 GB, 2 GB, 4 GB, 8 GB, 16 GB	1 GB, 2 GB, 4 GB, 8 GB	0.5 GB, 1 GB, 2 GB, 4 GB, 8 GB	0.5 GB, 1 GB, 2 GB, 4 GB, 8 GB	1 GB, 2 GB, 4 GB, 8 GB, 16 GB
Form Factor	PCIe 3.0 Dual Slot & SXM2 (rack servers)	PCIe 3.0 Dual Slot (rack servers)	PCIe 3.0 Single Slot (rack servers)	PCIe 3.0 Single Slot (rack servers)	PCIe 3.0 Dual Slot (rack servers)	PCIe 3.0 Dual Slot (rack servers)	MXM (blade servers)
Power	250W/300W	250W	70W	75W	300W (225W opt)	225W	90W
Thermal	passive	passive	passive	passive	active/passive	passive	bare board
	PERFORMANCE Optimized					DENSITY Optimized	BLADE Optimized

Figure 1 NVIDIA Tesla GPUs for virtualization workloads.

The NVIDIA Tesla T4 leverages the Turing architecture – the biggest architectural leap forward in over a decade – enabling major advances in efficiency and performance. Some of the key features provided by the NVIDIA Turing architecture include Tensor

Cores for accelerating deep learning inference workflows as well as CUDA cores, Tensor Cores, and RT Cores for real-time ray tracing acceleration and batch rendering. It's also the first GPU architecture to support GDDR6 memory, which provides improved performance and power efficiency versus the previous generation GDDR5.

The Tesla T4 is an RTX-capable GPU, benefiting from all of the enhancements of the RTX platform, including:

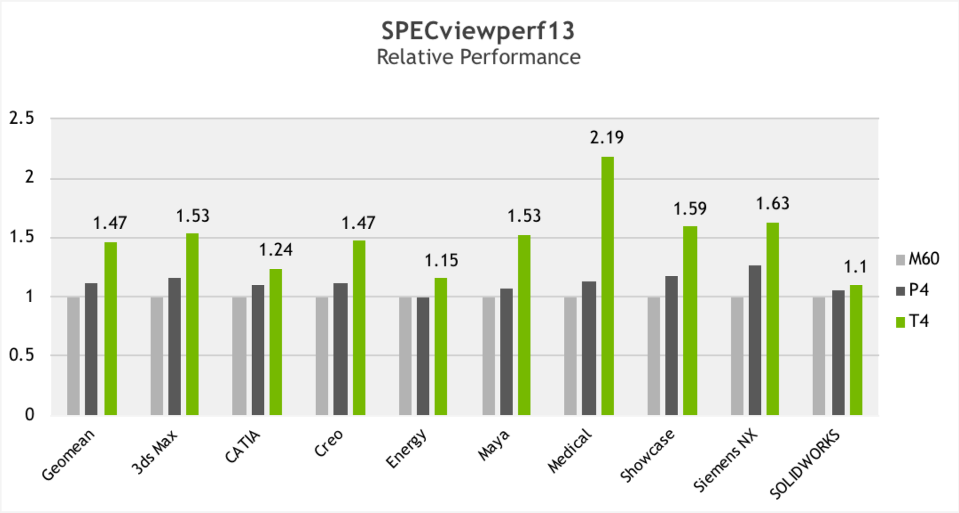
- Real-time ray tracing
- Accelerated batch rendering
- AI-enhanced denoising
- Photorealistic design with accurate shadows, reflections, and refractions

The T4 is well-suited for a wide range of data center workloads including:

- Virtual Desktops for knowledge workers using modern productivity applications
- Virtual Workstations for scientists, engineers, and creative professionals
- Deep Learning Inferencing and Training

HIGH-PERFORMANCE QUADRO VIRTUAL WORKSTATIONS

The graphics performance of the NVIDIA Tesla T4 directly benefits virtual workstations implemented with NVIDIA Quadro vDWS software to run rendering and simulation workloads.   Users of high end applications, such as CATIA, SOLIDWORKS and ArcGIS Pro, are typically segmented as light, medium or heavy based on the type of workflow they're running and the size of the model/data they are working with. The T4 is a low-profile, single slot card for light and medium users working with mid-to-large sized models. NVIDIA T4 offers double the amount of framebuffer (16GB) versus the previous generation P4 (8GB) card, therefore users can work with bigger models within their virtual workstations. Benchmark results show that T4 with Quadro vDWS delivers 25% faster performance than P4 and offers almost twice the professional graphics performance of the NVIDIA Tesla M60.



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

Figure 2 Tesla T4 performance comparison with M60 and P4 based on SPECviewperf13.

The Turing architecture of the Tesla T4 fuses real-time ray tracing, AI, simulation, and rasterization to fundamentally change computer graphics. Dedicated ray-tracing processors called RT Cores accelerate the computation of how light travels in 3D environments. Turing accelerates real-time ray tracing over the previous-generation NVIDIA Pascal™ architecture and can render final frames for film effects faster than CPUs. The new Tensor Cores, processors that accelerate deep learning training and inference, accelerate AI-enhanced graphics features—such as denoising, resolution scaling, and video re-timing—creating applications with powerful new capabilities.

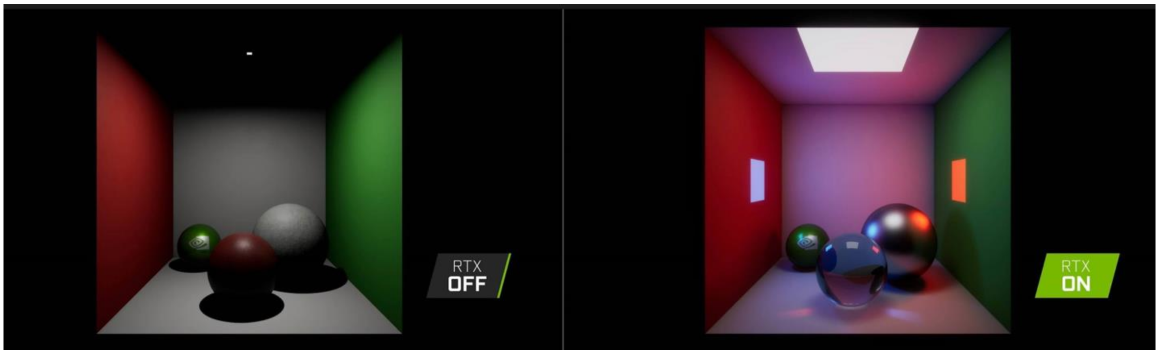


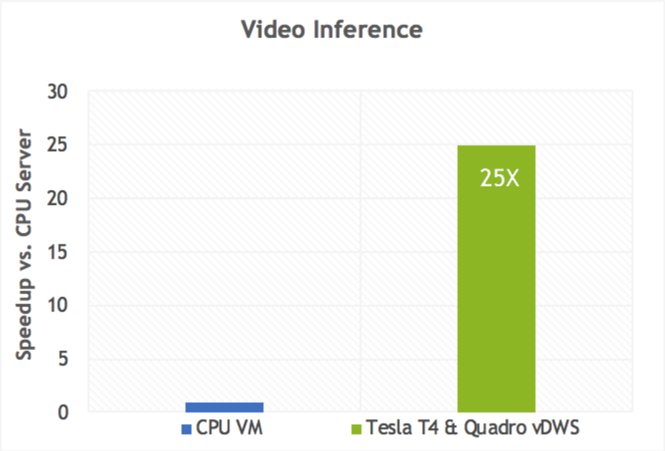
Figure 3 Illustrating the benefits of real-time rendering with RTX technology.

DEEP LEARNING INFERENCING

The NVIDIA T4 with the Turing architecture sets a new bar for power efficiency and performance for deep learning and AI. Its multi-precision tensor cores combined with accelerated containerized software stacks from NVIDIA GPU Cloud (NGC) delivers revolutionary performance.

As we are racing towards a future where every customer inquiry, every product and service will be touched and improved by AI, NVIDIA vGPU is bringing Deep Learning inferencing and training workflows to virtual machines. Quadro vDWS users can now execute inferencing workloads within their VDI sessions by accessing NGC containers. NGC integrates GPU-optimized deep learning frameworks, runtimes, libraries and even the OS into a ready-to-run container, available at no charge. NGC simplifies and standardizes deployment, making it easier and quicker for data scientists to build, train and deploy AI models. Accessing NGC containers within a VM offers even more portability and security to virtual users for classroom environments and virtual labs.

Test results show that Quadro vDWS users leveraging Tesla T4 can run deep learning inferencing workloads 25X faster than with CPU-only VMs.



Speedup: 25x faster

ResNet-50 (7ms latency limit)

Tested on a server with Intel Xeon Gold 6154 (18C, 3.0 GHz), Quadro vDWS with T4-16Q, VMware ESXi 6.7, host/guest driver 410.87/412.10, VM config, Ubuntu 16.04, 8 vCPU, 32GB memory. 25X performance improvement over CPU VM.

Figure 3 Run video inferencing workloads up to 25X faster with Tesla T4 and Quadro vDWS versus a CPU-only VM.

VIRTUAL DESKTOPS FOR KNOWLEDGE WORKERS

Benchmark test results show that the T4 is a universal GPU which can run a variety of workloads, including virtual desktops for knowledge workers accessing modern productivity applications. Modern productivity applications, high resolution and multiple monitors, and Windows 10 continue to require more graphics and with NVIDIA GRID vPC software, combined with NVIDIA Tesla GPUs, users can achieve a native-PC experience in a virtualized environment. While the Tesla M10 GPU, combined with GRID software, remains the ideal solution to provide optimal user density, TCO and performance for knowledge workers in a VDI environment, the versatility of the T4 makes it an attractive solution as well.

The M10 was announced in Spring of 2016 and offers the best user density and performance option for GRID vPC customers. The M10 is a 32GB dual slot card which draws up to 225W of power, therefore requires a supplemental power connector.   The Tesla T4 is a low profile, 16GB single slot card, which draws 70W maximum and does not require a supplemental power connector.

Two NVIDIA T4 GPUs provide 32GB of framebuffer and support the same user density as a single Tesla M10 with 32GB of framebuffer, but with lower power consumption. While the M10 provides the best value for knowledge worker deployments, selecting the T4 for this use case brings the unique benefits of the Turing architecture. This enables IT to maximize data center resources by running virtual desktops in addition to virtual workstations, deep learning inferencing, rendering and other graphics and compute intensive workloads -- all leveraging the same data center infrastructure. This ability to run mixed workloads can increase user productivity, maximize utilization, and reduce costs in the data center. Additional T4 technology enhancements include support for VP9 decode, which is often used for video playback, and H.265 (HEVC) 4:4:4 encode/decode.

SUMMARY

The flexible design of the Tesla T4 makes it well suited for any data center workload - enabling IT to leverage it for multiple use cases and maximize efficiency and utilization. It is perfectly aligned for vGPU implementations - delivering a native-PC experience for virtualized productivity applications, untethering architects, engineers and designers from their desks, and enabling deep learning inferencing workloads from anywhere, on any device. This universal GPU can be deployed across industry-standard servers to provide graphics and compute acceleration across any workload and future-proof the data center. Its dense, low power form factor can improve data center operating expenses while improving performance and efficiency and scales easily as compute and graphics needs grow.