NVIDIA T4 FOR VIRTUALIZATION

TB-09377-001-v01 | January 2019

Technical Brief

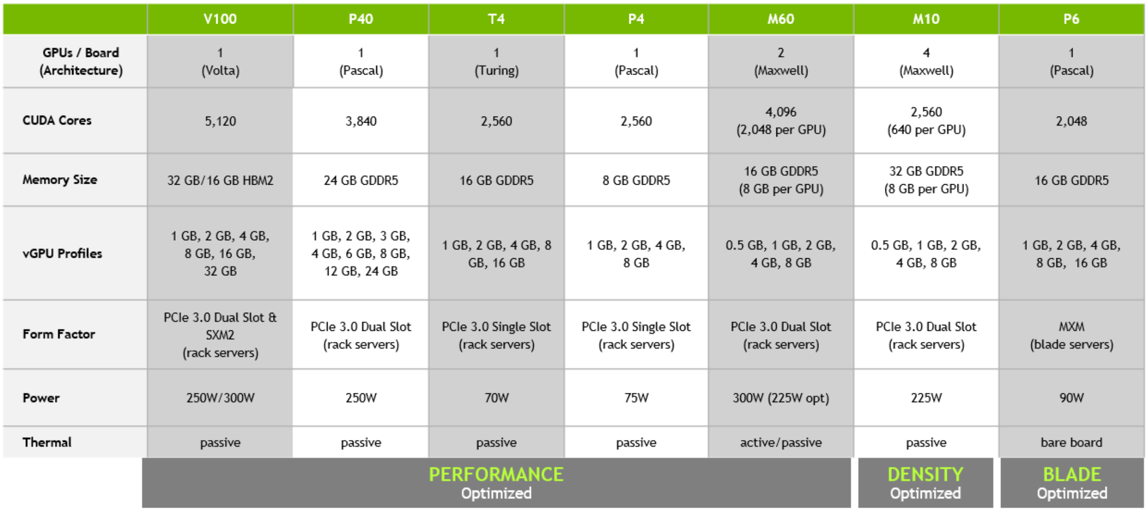
TABLE OF CONTENTS

POWERING ANY VIRTUAL WORKLOAD

The NVIDIA® T4 graphics processing unit (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 T4 is the most universal graphics processing unit (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 70 W 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 NVIDIA® Tesla® M10 or Tesla 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 GDDR5	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 T4 leverages the NVIDIA 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 NVIDIA® 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 T4 is an NVIDIA RTX™-capable GPU, benefiting from all of the enhancements of the NVIDIA RTX platform, including:

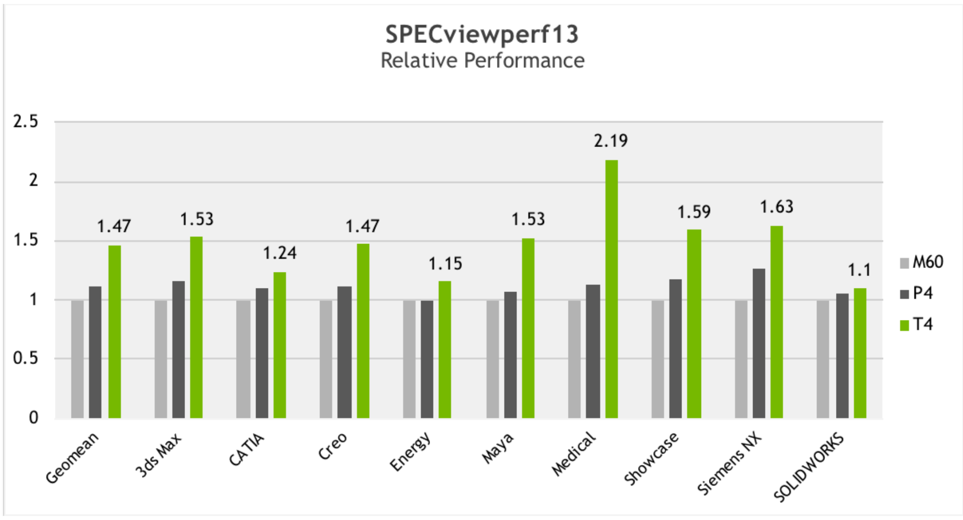
- Real-time ray tracing
- Accelerated batch rendering
- AI-enhanced denoising
- Photorealistic design with accurate shadows, reflections, and refractions

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 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. T4 offers double the amount of framebuffer (16 GB) versus the previous generation Tesla P4 (8 GB) 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 Tesla 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. T4 Performance Comparison with Tesla M60 and Tesla P4 Based on SPECviewperf13

The NVIDIA Turing architecture of the 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. NVIDIA 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.

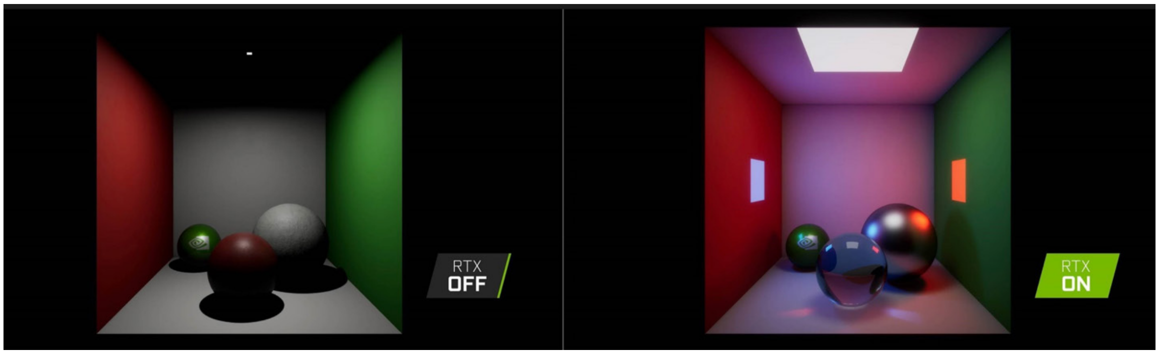


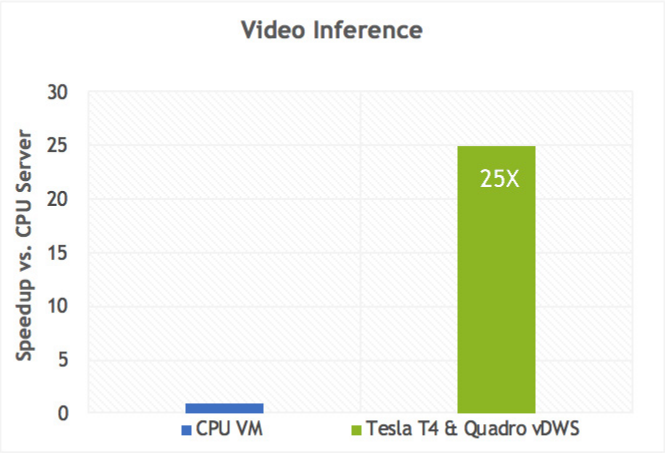
Figure 3. Benefits of Real-Time Rendering with NVIDIA RTX Technology

DEEP LEARNING INFERENCING

The T4 with the NVIDIA 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 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 4. Run Video Inferencing Workloads up to 25X Faster with T4 and Quadro vDWS vs. 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 NVIDIA 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 Tesla M10 was announced in Spring of 2016 and offers the best user density and performance option for NVIDIA GRID vPC customers. The Tesla M10 is a 32 GB dual-slot card which draws up to 225 W of power, therefore requires a supplemental power connector. The T4 is a low profile, 16 GB single-slot card, which draws 70 W maximum and does not require a supplemental power connector.

Two NVIDIA T4 GPUs provide 32 GB of framebuffer and support the same user density as a single Tesla M10 with 32 GB of framebuffer, but with lower power consumption. While the Tesla M10 provides the best value for knowledge worker deployments, selecting the T4 for this use case brings the unique benefits of the NVIDIA 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 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 on 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.

Notice

The information provided in this specification is believed to be accurate and reliable as of the date provided. However, NVIDIA Corporation (“NVIDIA”) does not give any representations or warranties, expressed or implied, as to the accuracy or completeness of such information. NVIDIA shall have no liability for the consequences or use of such information or for any infringement of patents or other rights of third parties that may result from its use. This publication supersedes and replaces all other specifications for the product that may have been previously supplied.

NVIDIA reserves the right to make corrections, modifications, enhancements, improvements, and other changes to this specification, at any time and/or to discontinue any product or service without notice. Customer should obtain the latest relevant specification before placing orders and should verify that such information is current and complete.

NVIDIA products are sold subject to the NVIDIA standard terms and conditions of sale supplied at the time of order acknowledgement, unless otherwise agreed in an individual sales agreement signed by authorized representatives of NVIDIA and customer. NVIDIA hereby expressly objects to applying any customer general terms and conditions with regard to the purchase of the NVIDIA product referenced in this specification.

NVIDIA products are not designed, authorized or warranted to be suitable for use in medical, military, aircraft, space or life support equipment, nor in applications where failure or malfunction of the NVIDIA product can reasonably be expected to result in personal injury, death or property or environmental damage. NVIDIA accepts no liability for inclusion and/or use of NVIDIA products in such equipment or applications and therefore such inclusion and/or use is at customer’s own risk.

NVIDIA makes no representation or warranty that products based on these specifications will be suitable for any specified use without further testing or modification. Testing of all parameters of each product is not necessarily performed by NVIDIA. It is customer’s sole responsibility to ensure the product is suitable and fit for the application planned by customer and to do the necessary testing for the application in order to avoid a default of the application or the product. Weaknesses in customer’s product designs may affect the quality and reliability of the NVIDIA product and may result in additional or different conditions and/or requirements beyond those contained in this specification. NVIDIA does not accept any liability related to any default, damage, costs or problem which may be based on or attributable to: (i) the use of the NVIDIA product in any manner that is contrary to this specification, or (ii) customer product designs.

No license, either expressed or implied, is granted under any NVIDIA patent right, copyright, or other NVIDIA intellectual property right under this specification. Information published by NVIDIA regarding third-party products or services does not constitute a license from NVIDIA to use such products or services or a warranty or endorsement thereof. Use of such information may require a license from a third party under the patents or other intellectual property rights of the third party, or a license from NVIDIA under the patents or other intellectual property rights of NVIDIA. Reproduction of information in this specification is permissible only if reproduction is approved by NVIDIA in writing, is reproduced without alteration, and is accompanied by all associated conditions, limitations, and notices.

ALL NVIDIA DESIGN SPECIFICATIONS, REFERENCE BOARDS, FILES, DRAWINGS, DIAGNOSTICS, LISTS, AND OTHER DOCUMENTS (TOGETHER AND SEPARATELY, “MATERIALS”) ARE BEING PROVIDED “AS IS.” NVIDIA MAKES NO WARRANTIES, EXPRESSED, IMPLIED, STATUTORY, OR OTHERWISE WITH RESPECT TO THE MATERIALS, AND EXPRESSLY DISCLAIMS ALL IMPLIED WARRANTIES OF NONINFRINGEMENT, MERCHANTABILITY, AND FITNESS FOR A PARTICULAR PURPOSE. Notwithstanding any damages that customer might incur for any reason whatsoever, NVIDIA’s aggregate and cumulative liability towards customer for the products described herein shall be limited in accordance with the NVIDIA terms and conditions of sale for the product.

Trademarks

Copyright