NVIDIA Tesla T4 Datasheet

Product Overview

- The NVIDIA® T4 belongs to the NVIDIA Turing™-based GPU series, designed for high-performance AI inference, deep learning, and virtualized workloads.
- As part of NVIDIA's AI computing ecosystem, the T4 is optimized for efficiency and scalability, making it an ideal choice for data centers, cloud computing, and edge AI applications.
- It delivers superior energy efficiency, leveraging Tensor Cores and RT Cores to accelerate AI-driven workloads, including speech recognition, image processing, and recommendation systems.
- Featuring the TU104 GPU architecture, the T4 is built with 16 GB of GDDR6 memory and operates with a maximum power consumption of just 70 W.
- Its single-slot, low-profile design allows for seamless deployment in dense server environments, while passive cooling ensures efficient thermal management with proper system airflow.
- The T4 supports mixed-precision computing, delivering exceptional performance across FP32, FP16, INT8, and INT4 workloads, making it well-suited for both training and inference tasks in modern AI applications.

Product Highlights

- **Revolutionary Turing Tensor Core Technology:** Accelerates diverse AI workloads with multi-precision computing.
- **Exceptional Performance:** Operates at just 70W, optimized for mainstream computing environments..
- **Energy-Efficient:** Utilizes a connector-to-power (C2P) airflow system with reverse configuration for optimal cooling.
- **High-Performance Video Transcoding:** Decodes up to 38 full-HD video streams with dedicated hardware.
- **Optimized for Cloud Workloads:** Ideal for high-performance computing, deep learning, machine learning, and data analytics

Detailed Features

Tesla T4

Component	Specification
Turing Tensor Cores	320
NVIDIA CUDA® cores	2,560
Capacity	16 GB GDDR6
Bandwidth	320+ GB/s
Power	70 watts

AI Video Transcoding Performance

Optimized for AI-driven video applications, the T4 supports up to 38 full-HD video streams simultaneously. Its dedicated hardware transcoding engines improve video analytics, surveillance, and real-time processing, benefiting industries like media, telecom, and security.

Multi-Precision Computing with Turing Tensor Cores

The T4 GPU leverages NVIDIA's Turing Tensor Cores for AI acceleration, dynamically switching between FP32, FP16, INT8, and INT4 precision. This flexibility enhances performance across deep learning training and inference, delivering up to 40X the AI performance of CPUs.

Hybrid Cloud Capabilities

Designed for hybrid cloud environments, the T4 enables seamless AI acceleration across on-premise and cloud infrastructures. It supports scalable AI deployment, allowing businesses to optimize performance while managing cloud resources efficiently.

Real-Time Inference Performance

With up to 40X faster inference than CPUs, the T4 ensures low-latency AI processing for applications like chatbots, recommendation systems, and visual search. Its high throughput enhances user experiences in customer service, e-commerce, and content delivery. Additionally, the T4 accelerates natural language processing (NLP) tasks, powering applications such as speech recognition, sentiment analysis, and automated translation with real-time responsiveness

Energy-Efficient AI Acceleration

Built for power efficiency, the T4 GPU delivers exceptional AI performance with just 70 W of power consumption. Its energy-efficient design significantly reduces operational costs for large-scale deployments, making it ideal for enterprises seeking sustainable AI solutions in data centers and edge computing environments.

Scalable Virtualization for Cloud Workloads

Supporting NVIDIA Virtual GPU technology, the T4 enhances cloud-based VDI (Virtual Desktop Infrastructure) and AI-driven virtual machines. It enables multiple users to access GPU resources concurrently, providing scalable performance for remote workstations, 3D rendering, and AI-powered cloud applications.

Enhanced Security and Data Protection

The T4 incorporates advanced security features, including secure boot and encrypted memory, to protect AI models and sensitive data. Its robust security framework ensures compliance with industry standards, making it suitable for mission-critical applications in healthcare, finance, and government sectors.

Advanced Graphics Capabilities

Beyond AI and video processing, the T4 excels in real-time rendering and graphics acceleration. With NVIDIA RTX technology, it supports ray tracing and high-performance visualization, making it a powerful solution for professionals in architecture, engineering, media production, and game development.

Technical Specifications

Product Specifications

Feature	Value
GPU SKU	TU104-895
GPU clocks	Base: 585 MHz, Maximum Boost: 1590 MHz
VBIOS EEPROM size	8 Mbit
UEFI Supported	Yes
PCI Express interface	PCI Express 3.0 ×16 x8
Thermal cooling solution	Passive

Feature	Value
Weight	301 Grams
Bracket Options	Full Height Bracket with screws (17 Grams) / Half Height Bracket with screws (10 Grams)
Maximum memory clock	5001 MHz
Memory size	16 GB
Memory bus width	256 bits
Peak Memory bandwidth	Up to 320 GB/s
SR-IOV support	Supported; 16 VF (virtual functions)
Operating temperature	0 °C to 50 °C
Storage temperature	40 °C to 75 °C
Operating humidity	5% to 90% relative humidity
Storage humidity	5% to 95% relative humidity
Mean time between failures (MTBF)	Uncontrolled environment: TBD at 35 °C, Controlled environment: TBD at 35 °C
Zero Power	Supported

Product Comparison

Feature	NVIDIA T4	NVIDIA V100	NVIDIA Tesla P4
GPU Architecture	Turing	Volta	Pascal
CUDA Cores	2,560	5,120	2,560
Tensor Cores	320	640	240
Base Clock	585 MHz	1,345 MHz	1,530 MHz
Boost Clock	1,590 MHz	1,530 MHz	1,530 MHz
Single Precision Performance (FP32)	8.1 TFLOPS	14 TFLOPS	5.4 TFLOPS
Mixed Precision (FP16/FP32)	65 FP16 TFLOPS	112 FP16 TFLOPS	21 FP16 TFLOPS
INT8 Performance	130 INT8 TOPS	200 INT8 TOPS	45 INT8 TOPS
INT4 Performance	260 INT4 TOPS	Not supported	Not supported

Feature	NVIDIA T4	NVIDIA V100	NVIDIA Tesla P4
Memory	16 GB GDDR6	16 GB HBM2	8 GB GDDR5
Memory Bandwidth	320 GB/s	900 GB/s	192 GB/s
Power Consumption	70 W	300 W	75 W
PCI Express Interface	PCIe Gen 3 x16	PCIe Gen 3 x16	PCIe Gen 3 x16
Form Factor	Low Profile PCIe	Full-Height PCIe	Low Profile PCIe
Cooling	Passive	Passive	Passive

Accessories

Category	Accessories
Included Accessories	- Power Cable
	- Mounting Bracket
	- Quick Installation Guide
Optional Accessories	-PCIe Riser Kit
	- 10Gb SFP+ Transceivers: 10GBASE-SR, 10GBASE-LR
	- Direct Attach Copper Cable (DAC): 10G, 25G
	- Active Optical Cables (AOC): 10G, 25G
	- Power Supply: 750W, 1300W High-Efficiency Power Supply
	- Cooling Fan Module: NVIDIA T4 Fan Module
	- Software Licenses: NVIDIA GPU Cloud (NGC) and CUDA Toolkit

Support & Warranty

24/7 Online Service

3-Year Premium Warranty

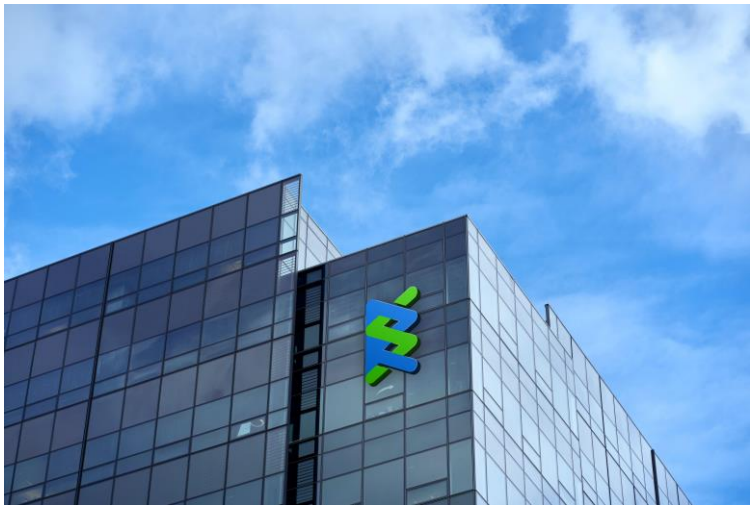
Professional Technical Support

100% Low Price Guarantee

100% Quality Assurance

100% Money Back Guarantee

About Us



Router-switch.com, headquartered in Hong Kong since 2002, has been a trusted global leader in ICT distribution for 23 years. We provide cutting-edge networking, cybersecurity, data center, and AI solutions to meet evolving business needs. Our wide range includes products from top brands like Cisco, Arista, Aruba, Fortinet, Mellanox, and Huawei, ensuring access to the latest technology and innovations.

18,000+ global customers	**600,000+** end-users	**200+** countries & regions	**23** years experience
500+ global vendors	**100,000+** SKUs available	**700+** local sales experts	**50-98%** off global list prices

Contact Us

Email

Sales Inquiries: sales@router-switch.com
Expert Technical Support: ccie-support@router-switch.com
Cooperative Partnerships: partner@router-switch.com

Phone

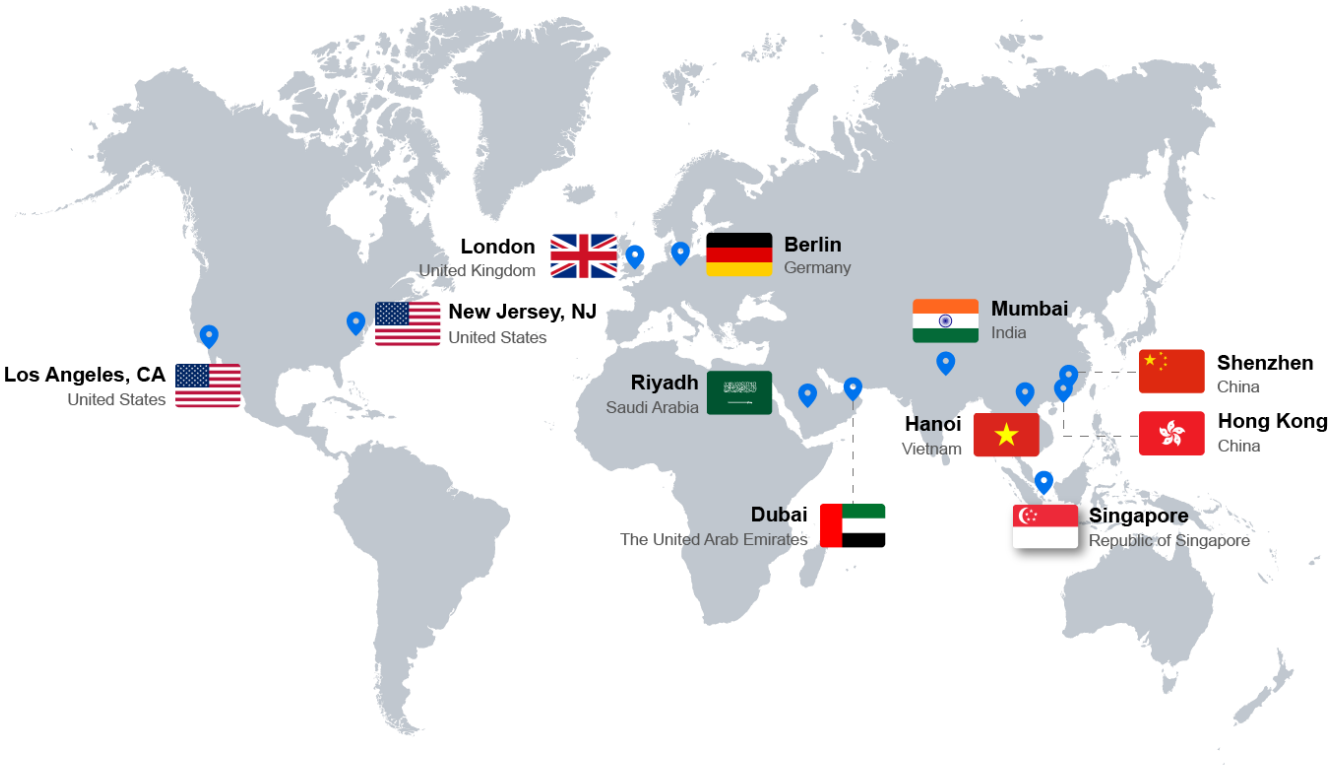
USA: +1-626-655-0998
Hong Kong: +852-25925389 / +852-25925411

Follow Us

Facebook: @Routerswitchdotcom
LinkedIn: Router-switch.com
X: @routerswitchcom
Instagram: @routerswitchdotcom

Global Footprint

Global Warehouses & Service Centers Across Continents.



Global Branches

Headquarters (Hong Kong)

Rm 605, 6/F, Fa Yuen Comm Bldg, 75-77 Fa Yuen St, Mongkok, Kowloon, Hong Kong, China

Shenzhen Branch

Jingfeng Building, 1001 Shangbu South Road, Futian District, Shenzhen, China

USA Branch

3592 Rosemead Blvd, B #220, Rosemead, CA 91770, USA

UK Branch

Third Floor, 207 Regent Street, London W1B 3HH, UK

References

[1] [1] NVIDIA. (n.d.). NVIDIA Tesla T4 GPU Accelerator. Retrieved from https://www.nvidia.com/en-us/data-center/tesla-t4/
[2] NVIDIA. (n.d.). NVIDIA Tesla T4 Tensor Core Product Brief. Retrieved from https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/tesla-t4/t4-tensor-core-product-brief.pdf