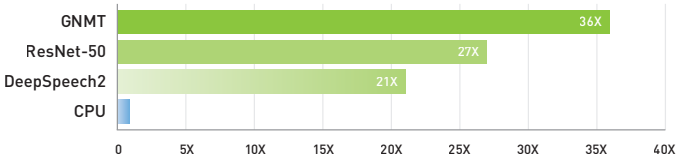
NVIDIA T4
TENSOR CORE GPU

GPU Acceleration Goes Mainstream

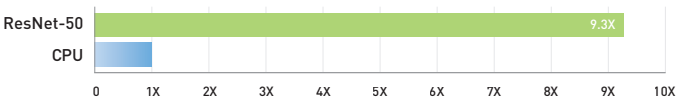
NVIDIA T4 enterprise GPUs supercharge the world's most trusted mainstream servers, easily fitting into standard data center infrastructures. Its low-profile, 70-watt (W) design is powered by NVIDIA Turing™ Tensor Cores, delivering revolutionary multi-precision performance to accelerate a wide range of modern applications, including machine learning, deep learning, and virtual desktops. This advanced GPU is packaged in an energy-efficient 70 W, small PCIe form factor, optimized for maximum utility in enterprise data centers and the cloud.

Inference Performance

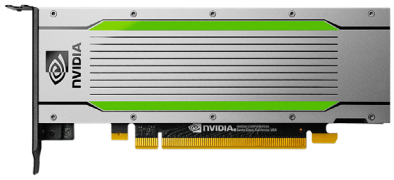


Comparisons made of one NVIDIA Tesla T4 GPU and servers with a dual-socket Xeon Gold 6140 CPU.

Training Performance



Comparison made between dual NVIDIA Tesla T4 GPUs and servers with a dual-socket Xeon Gold 6140 CPU.

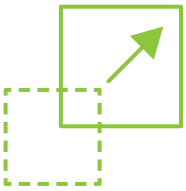


SPECIFICATIONS

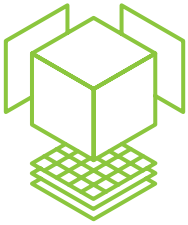
GPU Architecture	**NVIDIA Turing**
NVIDIA Turing Tensor Cores	**320**
NVIDIA CUDA® Cores	**2,560**
Single-Precision	**8.1 TFLOPS**
Mixed-Precision (FP16/FP32)	**65 TFLOPS**
INT8	**130 TOPS**
INT4	**260 TOPS**
GPU Memory	**16 GB GDDR6 300 GB/sec**
ECC	**Yes**
Interconnect Bandwidth	**32 GB/sec**
System Interface	**x16 PCIe Gen3**
Form Factor	**Low-Profile PCIe**
Thermal Solution	**Passive**
Compute APIs	**CUDA, NVIDIA TensorRT™, ONNX**

Performance to Drive Data Center Acceleration

The small-form-factor, 70-watt (W) design makes NVIDIA T4 optimized for scale-out servers, providing an incredible 50X higher energy efficiency compared to CPUs, drastically reducing operational costs. In the last two years, NVIDIA's inference platform has increased efficiency by over 10X, and it remains the most energy-efficient solution for distributed AI training and inference.



The NVIDIA T4 data center GPU is the ideal universal accelerator for distributed computing environments. Revolutionary multi-precision performance accelerates deep learning and machine learning training and inference, video transcoding, and virtual desktops. NVIDIA T4 supports all AI frameworks and network types, delivering dramatic performance and efficiency that maximize the utility of at-scale deployments.



Turing Tensor Core technology with multi-precision computing for AI powers breakthrough performance from FP32 to FP16 to INT8, as well as INT4 precisions. It delivers up to 9.3X higher performance than CPUs on training and up to 36X on inference.

Turing's powerful RT Cores, combined with NVIDIA RTX™ technology, enable real-time ray-traced rendering, delivering photorealistic objects and environments with physically accurate shadows, reflections, and refractions.

Arrow Contact Information

Email: IntelligentSolutions@arrow.com, **Online:** www.arrow.com/AIS/nvidia

To learn more about the NVIDIA T4, visit **www.nvidia.com/T4**