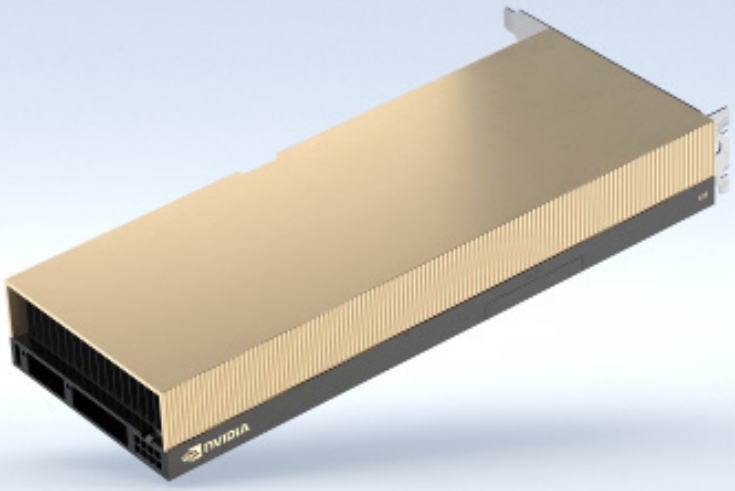
NVIDIA A30
TENSOR CORE GPU
VERSATILE COMPUTE ACCELERATION FOR MAINSTREAM ENTERPRISE SERVERS



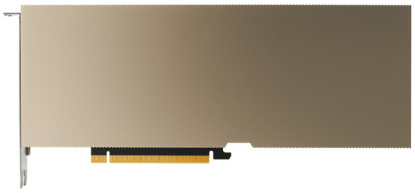
AI Inference and Mainstream Compute for Every Enterprise

NVIDIA A30 Tensor Core GPU is the most versatile mainstream compute GPU for AI inference and mainstream enterprise workloads. Powered by NVIDIA Ampere architecture Tensor Core technology, it supports a broad range of math precisions, providing a single accelerator to speed up every workload.

Built for AI inference at scale, the same compute resource can rapidly re-train AI models with TF32, as well as accelerate high-performance computing (HPC) applications using FP64 Tensor Cores. Multi-Instance GPU (MIG) and FP64 Tensor Cores combine with fast 933 gigabytes per second (GB/s) of memory bandwidth in a low 165W power envelope, all running on a PCIe card optimal for mainstream servers.

The combination of third-generation Tensor Cores and MIG delivers secure quality of service across diverse workloads, all powered by a versatile GPU enabling an elastic data center. A30's versatile compute capabilities across big and small workloads deliver maximum value for mainstream enterprises.

A30 is part of the complete NVIDIA data center solution that incorporates building blocks across hardware, networking, software, libraries, and optimized AI models and applications from NGC™. Representing the most powerful end-to-end AI and HPC platform for data centers, it allows researchers to deliver real-world results and deploy solutions into production at scale.



SYSTEM SPECIFICATIONS

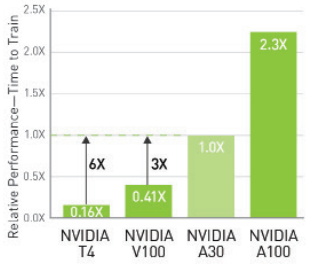
PNY Part Number	**NVA30TCGPU-KIT**
CUDA Cores	**3804**
Tensor Cores	**224**
Peak FP64	**5.2TF**
Peak FP64 Tensor Core	**10.3 TF**
Peak FP32	**10.3 TF**
TF32 Tensor Core	**82 TF | 165 TF***
BFLOAT16 Tensor Core	**165 TF | 330 TF***
Peak FP16 Tensor Core	**165 TF | 330 TF***
Peak INT8 Tensor Core	**330 TOPS | 661 TOPS***
Peak INT4 Tensor Core	**661 TOPS | 1321 TOPS***
Media engines	**1 optical flow accelerator (OFA) 1 JPEG decoder (NVJPEG) 4 Video decoders (NVDEC)**
GPU Memory	**24GB HBM2**
GPU Memory Bandwidth	**933GB/s**
Interconnect	**PCIe Gen4: 64GB/s Third-gen NVIDIA® NVLINK® 200GB/s****
Form Factor	**2-slot, full height, full length (FHFL)**
Max thermal design power (TDP)	**165W**
Multi-Instance GPU (MIG)	**4 MIGs @ 6GB each 2 MIGs @ 12GB each 1 MIGs @ 24GB**
Virtual GPU (vGPU) software support	**NVIDIA AI Enterprise for VMware NVIDIA Virtual Compute Server**

* With sparsity

Incredible Performance Across Workloads

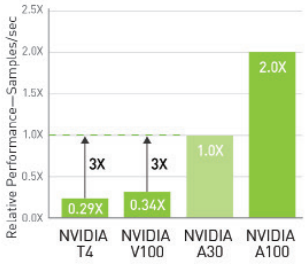
AI training—Up to 3X higher throughput than V100 and 6X higher than T4[1]

BERT Large Pre-Training (Normalized)



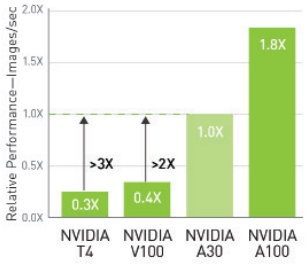
AI inference—Up to 3X higher throughput than V100 at real-time conversational AI[2]

BERT Large Inference (Normalized) Throughput for <10ms Latency



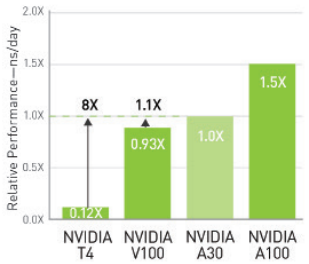
AI inference—Over 3X higher throughput than T4 at real-time image classification[3]

RN50 v1.5 Inference (Normalized) Throughput for <7ms Latency

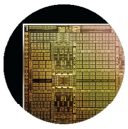


HPC—Up to 1.1X higher throughput than V100 and 8X higher than T4[4]

LAMMPS (Normalized)

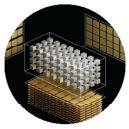


Groundbreaking Innovations



NVIDIA AMPERE ARCHITECTURE

Whether using MIG to partition an A30 GPU into smaller instances or NVIDIA NVLink to connect multiple GPUs to speed larger workloads, A30 can readily handle diverse-sized acceleration needs, from the smallest job to the biggest multi-node workload. A30 versatility means IT managers can maximize the utility of every GPU in their data center with mainstream servers, around the clock.

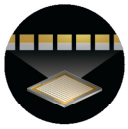


THIRD-GENERATION TENSOR CORES

NVIDIA A30 delivers 165 teraFLOPS (TFLOPS) of TF32 deep learning performance. That's 20X more AI training throughput and over 5X more inference performance compared to NVIDIA T4 Tensor Core GPU. For HPC, A30 delivers 10.3 TFLOPS of performance, nearly 30 percent more than NVIDIA V100 Tensor Core GPU.

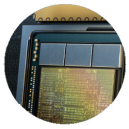
NEXT-GENERATION NVLINK

NVIDIA NVLink in A30 delivers 2X higher throughput compared to the previous generation. Two A30 PCIe GPUs can be connected via an NVLink Bridge to deliver 330 TFLOPs of deep learning performance.



MULTI-INSTANCE GPU (MIG)

An A30 GPU can be partitioned into as many as four GPU instances, fully isolated at the hardware level with their own high-bandwidth memory, cache, and compute cores. MIG gives developers access to breakthrough acceleration for all their applications. And IT administrators can offer right-sized GPU acceleration for every job, optimizing utilization and expanding access to every user and application.



HBM2

With up to 24GB of high-bandwidth memory (HBM2), A30 delivers 933GB/s of GPU memory bandwidth, optimal for diverse AI and HPC workloads in mainstream servers.

STRUCTURAL SPARSITY

AI networks have millions to billions of parameters. Not all of these parameters are needed for accurate predictions, and some can be converted to zeros, making the models "sparse" without compromising accuracy. Tensor Cores in A30 can provide up to 2X higher performance for sparse models. While the sparsity feature more readily benefits AI inference, it can also improve the performance of model training.

The End-to-End Solution for Enterprises

NVIDIA A30 Tensor Core GPU— powered by the NVIDIA Ampere architecture, the heart of the modern data center—is an integral part of the NVIDIA data center platform. Built for deep learning, HPC, and data analytics, the platform accelerates over 2,000 applications, including every major deep learning framework. Additionally, NVIDIA AI Enterprise, an end-to-end, cloud-native suite of AI and data analytics software, is certified to run on A30 in hypervisor-based virtual infrastructure with VMware vSphere. This enables management and scaling of AI workloads in a hybrid cloud environment. The complete NVIDIA platform is available everywhere, from data center to edge, delivering both dramatic performance gains and cost-saving opportunities.

OPTIMIZED SOFTWARE AND SERVICES FOR ENTERPRISE

NVIDIA CERTIFIED

EVERY DEEP LEARNING FRAMEWORK

mxnet PYTORCH APACHE Spark TensorFlow

2,000+ GPU-ACCELERATED APPLICATIONS

To learn more about the NVIDIA A30 Tensor Core GPU, visit **www.pny.com/a30**

[1] BERT-Large Pre-Training (9/10 epochs) Phase 1 and (1/10 epochs) Phase 2, Sequence Length for Phase 1 = 128 and Phase 2 = 512, dataset = real, NGC™ container = 21.03,8x GPU: T4 (FP32, BS=8, 2) | V100 PCIE 16GB (FP32, BS=8, 2) | A30 (TF32, BS=8, 2) | A100 PCIE 40GB (TF32, BS=54, 8) | batch sizes indicated are for Phase 1 and Phase 2 respectively

[2] NVIDIA® TensorRT®, Precision = INT8, Sequence Length = 384, NGC Container 20.12, Latency <10ms, Dataset = Synthetic; 1x GPU: A100 PCIe 40GB (BS=8) | A30 (BS=4) | V100 SXM2 16GB (BS=1) | T4 (BS=1)

[3] TensorRT, NGC Container 20.12, Latency <7ms, Dataset=Synthetic,; 1x GPU: T4 (BS=31, INT8) | V100 (BS=43, Mixed precision) | A30 (BS=96, INT8) | A100 (BS=174, INT8)

[4] Dataset: ReaxFF/C, FP64 | 4x GPU: T4, V100 PCIE 16GB, A30

© 2021 NVIDIA Corporation. All rights reserved. NVIDIA, the NVIDIA logo, NGC, and NVLink are trademarks and/or registered trademarks of NVIDIA Corporation in the U.S. and other countries. All other trademarks and copyrights are the property of their respective owners. APR21

PNY | NVIDIA