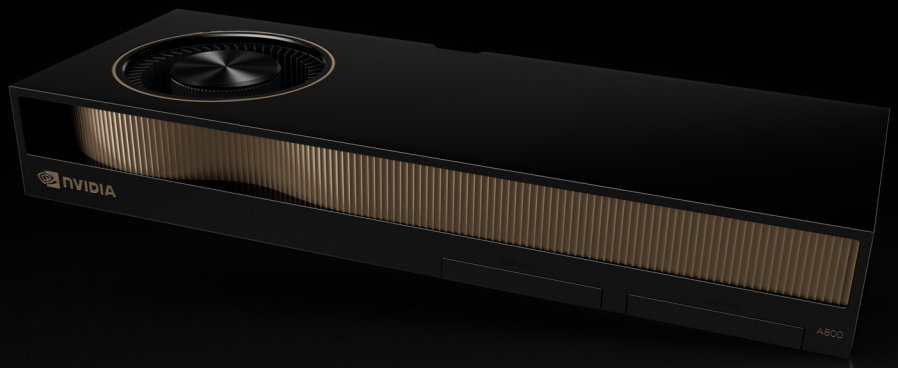
Datasheet

NVIDIA A800 40GB Active



The ultimate workstation development platform for AI, data science, and high-performance computing (HPC).

Breakthrough Performance for Advanced Computing Workloads

Bring the power of a supercomputer to your workstation and accelerate end-to-end data science workflows with the NVIDIA A800 40GB Active GPU. Powered by the NVIDIA Ampere architecture, the A800 40GB Active delivers powerful compute, high-speed memory, and scalability, so data professionals can tackle their most challenging data science, AI, and HPC workloads:

Data Science and Analytics

Power complex data science workflows and accelerate the end-to-end data science pipeline, from data loading and data manipulation to machine learning and visualization.

AI Training and Inference

Conquer demanding AI development, training, and inference workflows, including data preparation and processing, model optimization and tuning, and early-stage training.

HPC

Run large-scale simulations in full FP64 precision with incredible speed, shortening development timelines and accelerating time to value.

Supercharge AI Development Out of the Box With NVIDIA AI Enterprise

Each NVIDIA A800 40GB Active GPU comes with a three-year subscription to NVIDIA AI Enterprise, an end-to-end software platform with enterprise security, stability, manageability, and support. NVIDIA AI Enterprise includes 100+ AI frameworks, libraries, pretrained models, and tools for rapid development and deployment of production-ready AI and data science. Together with NVIDIA A800 40GB, NVIDIA AI Enterprise simplifies AI adoption and achieves business insights faster with the highest performance. Access the NVIDIA AI Enterprise software subscription and learn more about its benefits.

Key Features

NVIDIA Ampere Architecture

Third-Generation Tensor Cores

- Powerful double-precision (FP64) capabilities
- Accelerated training and inference performance

Third-Generation NVIDIA® NVLink™

- Connect two A800 GPUs to scale up to 80 gigabytes (GB) of memory
- 400 gigabytes per second (GB/s) of bidirectional bandwidth

Ultra-Fast HBM2 Memory

- 40GB of high-speed HBM2 memory
- 1.5 TB/s of memory bandwidth

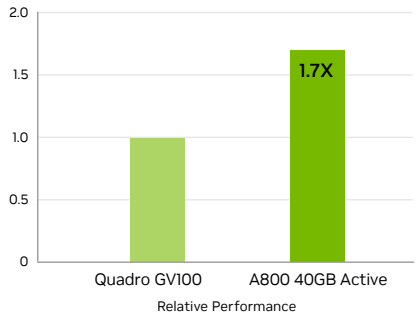
Multi-Instance GPU (MIG)

- Fully isolated and secure multi-tenancy
- Partition up to seven instances

Specifications	
GPU Memory	40GB HBM2
Memory Interface	5,120-bit
Memory Bandwidth	1.5 TB/s
CUDA® Cores	6,912
Tensor Cores	432
Double-Precision Performance	9.7 TFLOPS
Single-Precision Performance	19.5 TFLOPS
Peak Tensor Performance	623.8 TFLOPS
Multi-Instance GPU	Up to 7 MIG instances @ 5GB
NVIDIA NVLink	Yes
NVLink Bandwidth	400GB/s
Graphics Bus	PCIe 4.0 x 16
Power Consumption	240W
Thermal	Active
Form Factor	4.4” H x 10.5” L, dual slot
Display Capability[1]	-

HPC

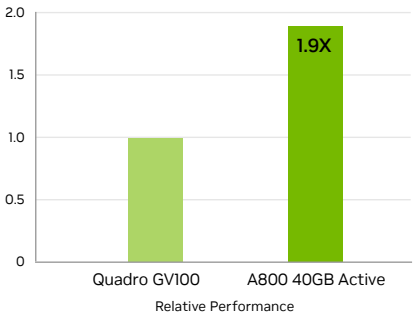
LAMMPS



Tests run on an Intel Xeon Gold 6126 processor, NVIDIA Driver 535.104. Relative speedup for LAMMPS patch_8Feb2023, Atomic Fluid Lennard-Jones 2.5 (cutoff); Precision=FP64;

HPC

GTC



Tests run on an Intel Xeon Gold 6126 processor, NVIDIA Driver 535.104. Relative speedup for GTC Version 4.5, TAE, Precision=FP32

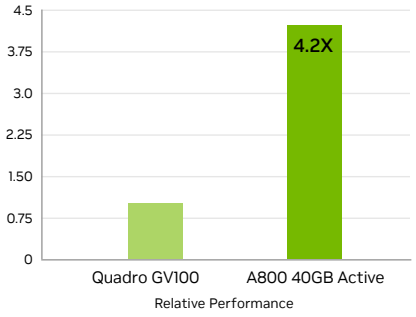
NVIDIA AI Enterprise Software

An End-to-End AI Software Platform

- A three-year NVIDIA AI Enterprise license included with each A800 40GB Active GPU
- Fast time to production for AI with access to AI frameworks, libraries, and tools
- Enterprise security, stability, manageability, and support
- Software activation required

AI Inference

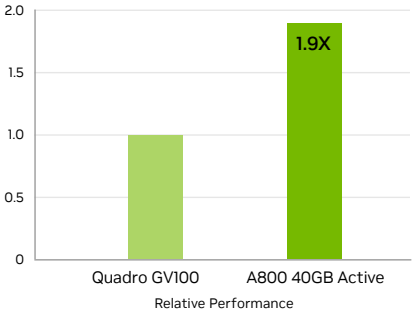
BERT - Large



Tests run on an Intel Xeon Gold 6126 processor, NVIDIA Driver 535.104. Relative speedup for GPT2 Inference . Batch Size=32; Precision=Mixed; Data=Synthetic; cuDNN Version=8.9.3.22;

AI Training

BERT - Large



Tests run on an Intel Xeon Gold 6126 processor, NVIDIA Driver 535.104. Relative speedup for BERT Large Pre-Training Phase 2 Batch Size=8; Precision=Mixed; AMP=Yes; Data=Real; Sequence Length=512; Gradient Accumulation Steps=_SEE_OUTPUTS_; cuDNN Version=8.9.3.28; NCCL Version=2.18.3

Ready to Get Started?

To learn more about NVIDIA A800 40GB Active, visit www.nvidia.com/a800

To activate your 3-year subscription of NVIDIA AI Enterprise, visit www.nvidia.com/activate-license

1. The A800 40GB Active does not come equipped with display ports. The NVIDIA RTX 4000 Ada Generation, NVIDIA RTX A4000, and the NVIDIA T1000 have been qualified to support display out capabilities.

© 2023 NVIDIA Corporation. All rights reserved. NVIDIA, the NVIDIA logo, CUDA, NVLink, Quadro, and RTX are trademarks and/or registered trademarks of NVIDIA Corporation in the U.S. and other countries. Other company and product names may be trademarks of the respective companies with which they are associated. All other trademarks are the property of their respective owners. 2819988. NOV23

PNY Technologies Europe
Rue Joseph Cugnot BP 40181
33708 Mérignac Cedex | France T +33 (0)5 56 13 75 75 | F +33 (0)5 56 13 75 77
For more information visit: www.pny.eu