NVIDIA Tensor Cores

Unprecedented Acceleration for Generative AI

Tensor Cores enable mixed-precision computing, dynamically adapting calculations to accelerate throughput while preserving accuracy and providing enhanced security. The latest generation of Tensor Cores are faster than ever on a broad array of AI and high-performance computing (HPC) tasks. From 4X speedups in training trillion-parameter generative AI models to a 30X increase in inference performance, NVIDIA Tensor Cores accelerate all workloads for modern AI factories.

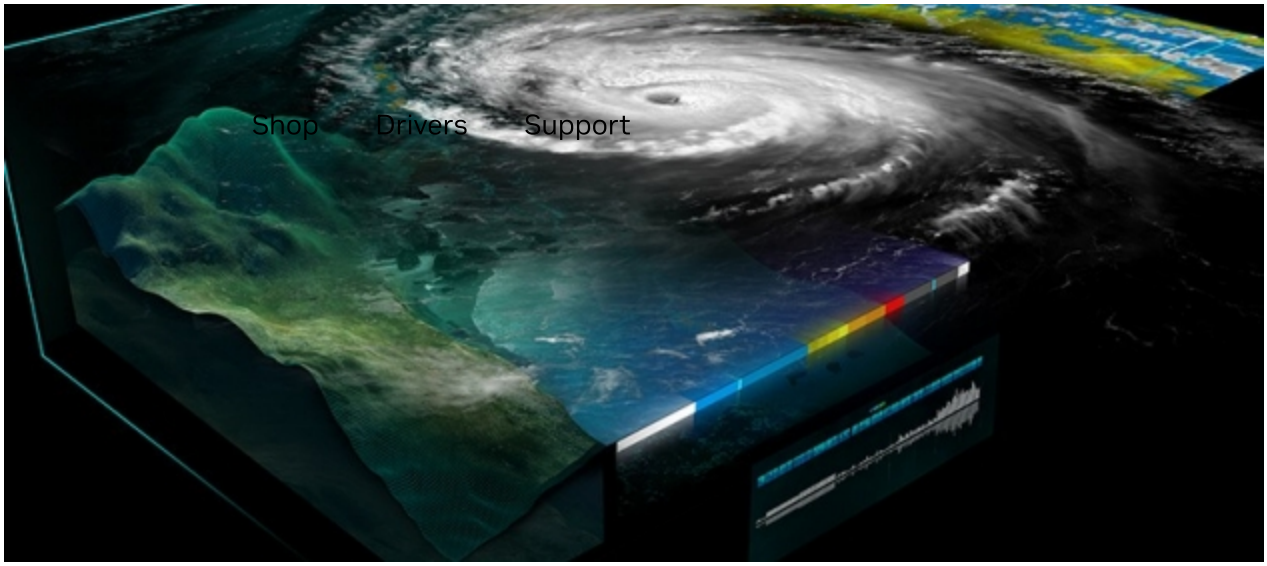
Introduction Blackwell Hopper Specifications

Training multi-trillion-parameter generative AI models in 16-bit floating point (FP16) precision can take months. NVIDIA Tensor Cores provide an order-of-magnitude higher performance with reduced precisions like FP8 in the Transformer Engine. With direct support in native frameworks via CUDA-X™ libraries, implementation is automatic, which dramatically slashes training-to-convergence times while maintaining accuracy.

Breakthrough Inference

Achieving low latency at high throughput while maximizing utilization is the most important performance requirement of deploying inference reliably. The NVIDIA Blackwell architecture's second-generation Transformer Engine delivers exceptional performance and also has the versatility to accelerate diverse multi-trillion-parameter generative AI models.

Tensor Cores has enabled NVIDIA to win MLPerf industry-wide benchmarks for inference.



Advanced HPC

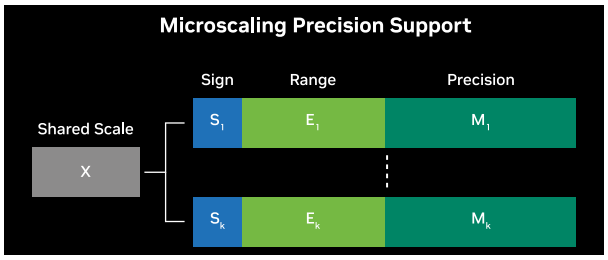
HPC is a fundamental pillar of modern science. To unlock next-generation discoveries, scientists use simulations to better understand complex molecules for drug discovery, physics for potential sources of energy, and atmospheric data to better predict and prepare for extreme weather patterns. NVIDIA Tensor Cores offer a full range of precisions, including FP64, to accelerate scientific computing with the highest accuracy needed.

The HPC SDK provides the essential compilers, libraries, and tools for developing HPC applications for the NVIDIA platform.

NVIDIA Blackwell Tensor Cores

Fifth Generation

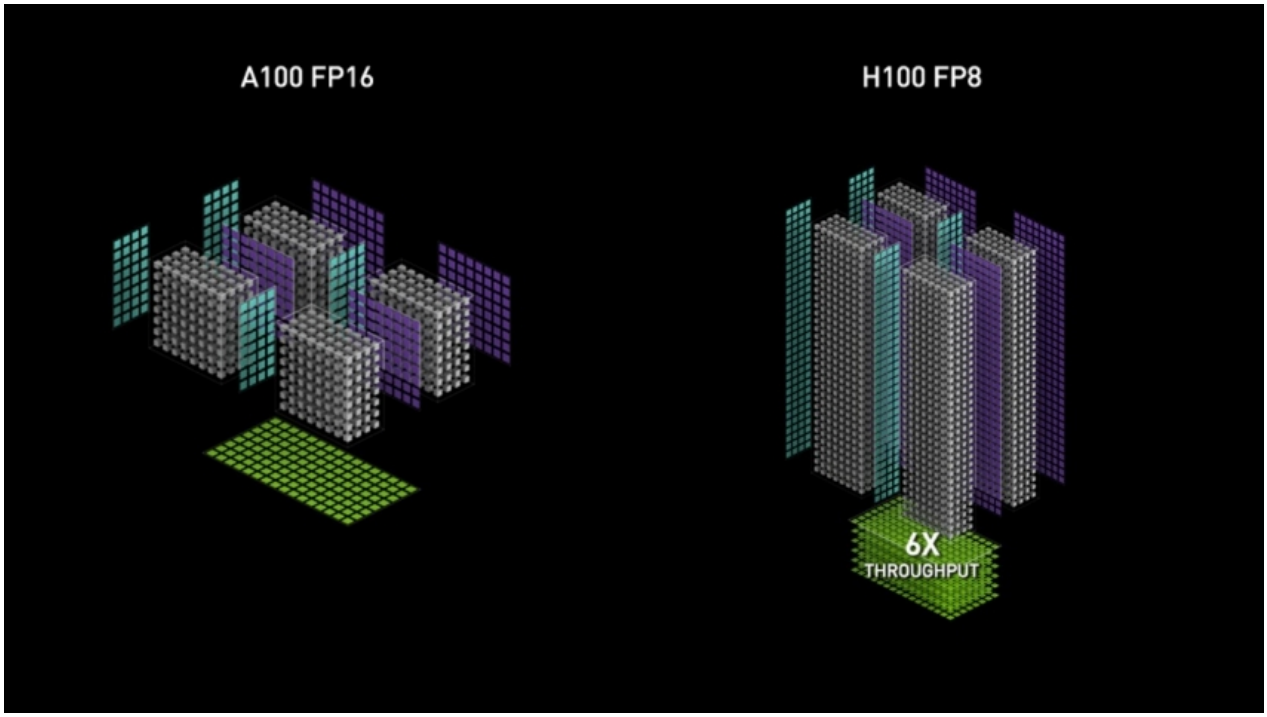
The Blackwell architecture delivers a 30X speedup compared to the previous NVIDIA Hopper™ generation for massive models such as GPT-MoE-1.8T. This performance boost is made possible with the fifth-generation of Tensor Cores. Blackwell Tensor Cores add new precisions, including community-defined microscaling formats, giving better accuracy and



New Precision Formats

As generative AI models explode in size and complexity, it's critical to improve training and inference performance. To meet these compute needs, Blackwell Tensor Cores support new quantization formats and precisions, including community-defined microscaling formats.

NVIDIA Hopper Architecture Tensor Cores



Fourth Generation

Transformer Engine, using FP8 to deliver 6X higher performance over FP16 for trillion-parameter-model training. Combined with 3X more performance using TF32, FP64, FP16, and INT8 precisions, Hopper Tensor Cores deliver speedups to all workloads.

Learn More About the NVIDIA Hopper Architecture ›

The Most Powerful End-to-End AI and HPC Data Center Platform

Tensor Cores are essential building blocks of the complete NVIDIA data center solution that incorporates hardware, networking, software, libraries, and optimized AI models and applications from the NVIDIA NGC™ catalog. The most powerful end-to-end AI and HPC platform, it allows researchers to deliver real-world results and deploy solutions into production at scale.

	Blackwell	Hopper
Supported Tensor Core precisions	FP64, TF32, BF16, FP16, FP8, INT8, FP6, FP4	FP64, TF32, BF16, FP16, FP8, INT8
Supported CUDA® Core precisions	FP64, FP32, FP16, BF16	FP64, FP32, FP16, BF16, INT8

*Preliminary specifications, may be subject to change

Learn More About NVIDIA Blackwell.

Products

Shop Drivers Support

Data Center GPUs

NVIDIA DGX Platform

NVIDIA EGX Platform

NVIDIA HGX Platform

Networking Products

Virtual GPUs

Technologies

NVIDIA Blackwell Architecture

NVIDIA Hopper Architecture

Confidential Computing

NVLink-C2C

NVLink/NVSwitch

Tensor Cores

Multi-Instance GPU

IndeX ParaView Plugin

Cybersecurity - Morpheus

Data Center Blogs

Data Center GPUs Product Literature

DGX Product Literature

Documentation

Energy Efficiency Calculator

Glossary

GPU Apps Catalog

GPU Test Drive

GTC AI Conference

NVIDIA GRID Community Advisors

Qualified System Catalog

Technical Blog

Technical Training

Training for IT Professionals

Where to Buy

Virtual GPU Forum

Virtual GPU Product Literature

Shop Drivers Support

Company Overview

Investors

Venture Capital (NVentures)

NVIDIA Foundation

Research

Social Responsibility

Technologies

Careers

Follow Data Center

United States





Shop Drivers Support