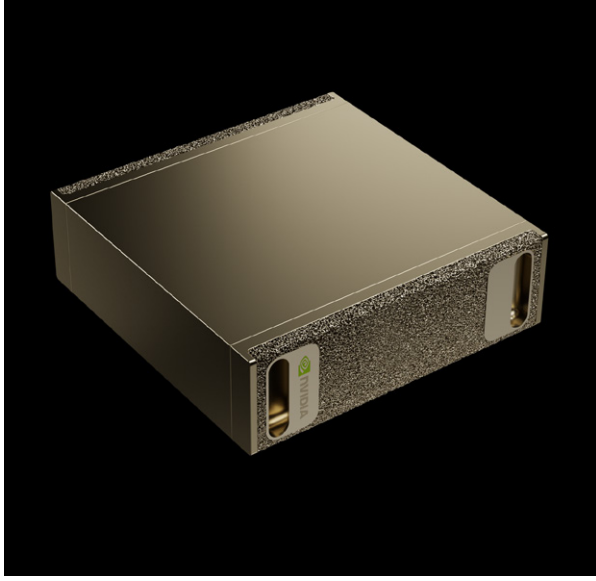
Datasheet

NVIDIA DGX Spark

DGX personal AI computer, designed to build and run AI.



Desktop AI Compute Demands

The increasing size and complexity of generative AI models is making development efforts on local systems challenging. Prototyping, tuning, and inferencing large models locally requires large amounts of memory and significant compute performance. As enterprises, software providers, government agencies, startups, and researchers staff up AI efforts, the need for AI compute resources continues to grow.

200B Parameter Models on Your Desk

NVIDIA DGX™ Spark is part of a new class of computers designed from the ground up to build and run AI. Powered by the NVIDIA GB10 Grace Blackwell Superchip and based on the NVIDIA Grace Blackwell architecture, NVIDIA DGX Spark delivers up to 1 petaFLOP[1] of AI performance to power large AI workloads. With 128 GB of unified system memory, developers can experiment, fine-tune, or inference models of up to 200B parameters.[2] Plus, NVIDIA ConnectX™ networking can connect two NVIDIA DGX Spark supercomputers to enable inference on models up to 405B parameters.[2]

To give developers a familiar experience, NVIDIA DGX Spark mirrors the same software architecture that powers industrial-strength AI factories. Using the NVIDIA DGX OS with Ubuntu Linux and preconfigured with the latest NVIDIA AI software stack, along with developer program access to NVIDIA NIM™ and NVIDIA Blueprints, developers can hit the ground running using common tools such as PyTorch, Jupyter, and Ollama to prototype, fine-tune, and inference on NVIDIA DGX Spark and easily move work to the data center or cloud.

By delivering incredible performance and capabilities in a compact package, NVIDIA DGX Spark lets developers, researchers, data scientists, and students continue to push the boundaries of generative AI.

Built on NVIDIA Grace Blackwell

At the heart of NVIDIA DGX Spark is the NVIDIA GB10 Grace Blackwell Superchip based on the NVIDIA Grace Blackwell architecture optimized for a desktop form factor. GB10 features a powerful NVIDIA Blackwell GPU with fifth-generation Tensor Cores and FP4 support, delivering up to 1 petaFLOP[1] of AI compute. GB10 includes a high-performance Grace 20-core Arm CPU to supercharge data preprocessing and orchestration, speeding up model tuning and inferencing. The GB10 Superchip uses the NVIDIA NVLink™-C2C to deliver a CPU+GPU coherent memory model with 5x the bandwidth of PCIe Gen 5.

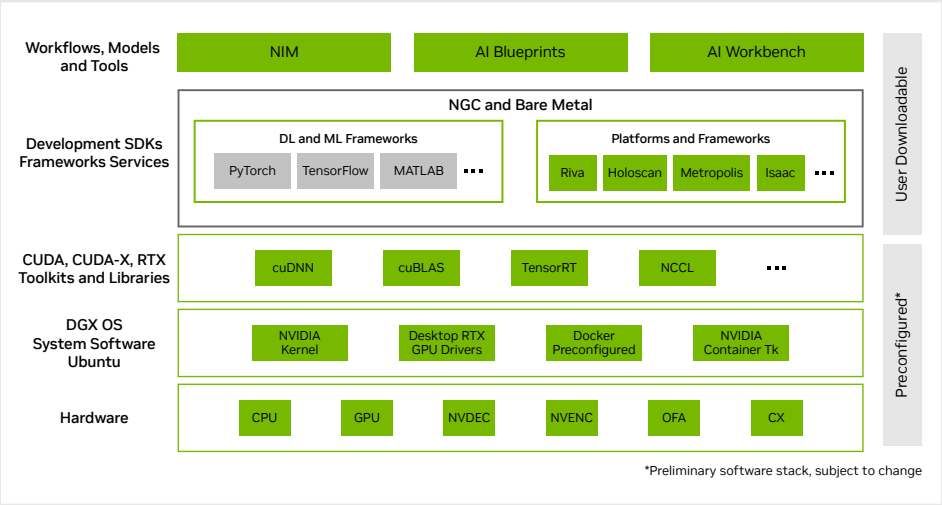
Key Features

- Built on NVIDIA GB10 Grace Blackwell Superchip
- NVIDIA Blackwell GPU with fifth-generation Tensor Core technology
- NVIDIA Grace CPU with 20-core high-performance Arm architecture
- Up to 1 petaFLOP[1] of AI performance using FP4
- 128 GB of coherent, unified system memory
- Support for up to 200 billion parameter[2] models
- NVIDIA ConnectX™ networking to link two systems to work with models up to 405 billion parameters
- 4 TB of NVMe storage
- Compact desktop form factor

Work With Large-Parameter AI Models

With 128 GB of unified system memory and support for the FP4 data format, NVIDIA DGX Spark can support AI models of up to 200B parameters[2], enabling AI developers to prototype, fine-tune, and inference large models on their desktop. With built-in NVIDIA ConnectX network technology, two NVIDIA DGX Spark systems can be connected to work on even larger models such as Llama 3.1 405B.

Develop Locally, Deploy Anywhere at Scale



NVIDIA DGX Spark software stack

NVIDIA DGX Spark provides organizations and developers with a powerful, economical experimentation ground for prototype models, freeing up valuable compute resources in their cluster environments better suited for training and deploying production models. Leveraging the NVIDIA AI platform software architecture makes it possible for NVIDIA DGX Spark users to easily move their work from their desktop to DGX Cloud or any accelerated cloud or data center infrastructure, making it easier than ever to prototype, fine-tune, and iterate.

Technical Specifications*	
Architecture	NVIDIA Grace Blackwell
GPU	NVIDIA Blackwell Architecture
CPU	20 core Arm, 10 Cortex-X925 + 10 Cortex-A725 Arm
CUDA Cores	NVIDIA Blackwell Generation
Tensor Cores	5th Generation
RT Cores	4th Generation
Tensor Performance[1]	1 PFLOP
System Memory	128 GB LPDDR5x, coherent unified system memory
Memory Interface	256-bit
Memory Bandwidth	Up to 273 GB/s
Storage	4 TB NVME.M2 with self-encryption
USB	4x USB TypeC
Ethernet	1x RJ-45 connector 10 GbE
NIC	ConnectX-7 NIC @ 200 Gbps
Wi-Fi	WiFi 7
Bluetooth	BT 5.4 w/LE
Audio-output	HDMI multichannel audio output
Power Consumption	240 W
Display Connectors	1x HDMI 2.1a
NVENC | NVDEC	1x | 1x
OS	NVIDIA DGX™ OS
System Dimensions	150 mm L x 150 mm W x 50.5 mm H
System Weight	1.2 kg

* preliminary specifications, subject to change

Ready to Get Started?

To learn more about NVIDIA DGX Spark, visit
www.openzeka.com/dgx-spark/

1. Theoretical FP4 TOPS using the sparsity feature.
2. Using FP4 precision models.

© 2025 NVIDIA Corporation and affiliates. All rights reserved. NVIDIA, the NVIDIA logo, ConnectX, DGX, NVLink, and NIM are trademarks and/or registered trademarks of NVIDIA Corporation and affiliates in the U.S. and other countries. Other company and product names may be trademarks of the respective owners with which they are associated. 4319150SEP25

OPEN ZEKA

NVIDIA