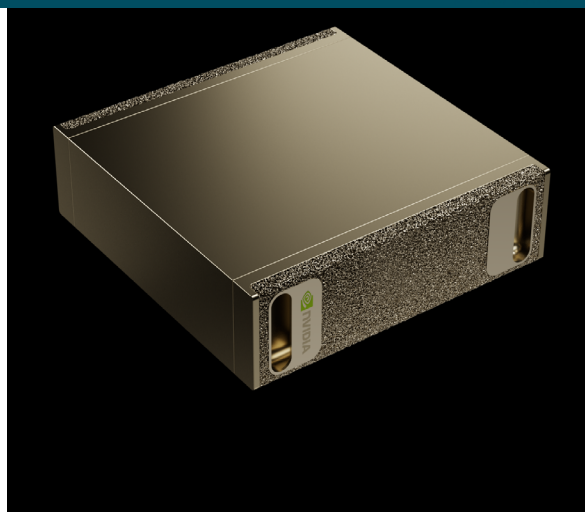
数据表

NVIDIA

NVIDIA DGX Spark

DGX 个人 AI 计算机，专为构建和运行 AI 而设计。



桌面端 AI 计算需求

生成式 AI 模型的规模和复杂性与日俱增，这使得本地系统的开发工作更具挑战性。在本地进行大模型的原型设计、微调和推理需要大量显存和出色的计算性能。随着企业、软件提供商、政府机构、初创企业和研究人员不断加大 AI 开发工作力度，对 AI 计算资源的需求也持续增长。

桌面端的 200B 参数模型

NVIDIA DGX™ Spark 是专为完整构建和运行 AI 而设计的新型计算机。NVIDIA DGX Spark 搭载了 NVIDIA GB10 Grace Blackwell 超级芯片，基于先进的 NVIDIA Grace Blackwell 架构设计，能够提供高达 1 petaFLOP 的性能，为大型 AI 工作负载提供动力。借助 128 GB 的统一寻址系统内存，开发人员可以对多达 200B 参数的模型进行试验、微调或推理。此外，NVIDIA ConnectX™ 网络可以连接两台 NVIDIA DGX Spark AI 超级计算机，从而支持对多达 405B 参数的模型进行推理。

为了给开发者带来熟悉的体验，NVIDIA DGX Spark 采用了与工业级 AI 工厂相同的软件架构。基于 NVIDIA DGX OS 和 Ubuntu Linux 并预先配置最新的 NVIDIA AI 软件堆栈，以及对开发人员开放 NVIDIA NIM™ 和 NVIDIA Blueprint 的访问权限，开发人员可以使用 Pytorch、Jupyter 和 Ollama 等常用工具在 NVIDIA DGX Spark 上进行原型设计、微调和推理，并轻松将任务迁移至数据中心或云端。

NVIDIA DGX Spark 在小巧的封装中提供出色的性能和强大的功能，助力开发者、研究人员、数据科学家和学生继续突破生成式 AI 的边界。

基于 NVIDIA Grace Blackwell 架构

NVIDIA DGX Spark 的核心是 GB10 Grace Blackwell 超级芯片，该芯片基于 NVIDIA Grace Blackwell 架构，并针对桌面端外形进行了优化。GB10 配备功能强大的 NVIDIA Blackwell GPU，支持第五代 Tensor Core 和 FP4，可提供高达 1 petaFLOP[1] 的 AI 计算性能。GB10 还包含高性能 Grace 20-core Arm CPU，可强效助力数据预处理和编排，从而加速模型调整和实时推理。GB10 超级芯片使用 NVLink™-C2C 互联技术，提供 CPU + GPU 相结合的一致性内存模型，带宽是第五代 PCIe 的 5 倍。

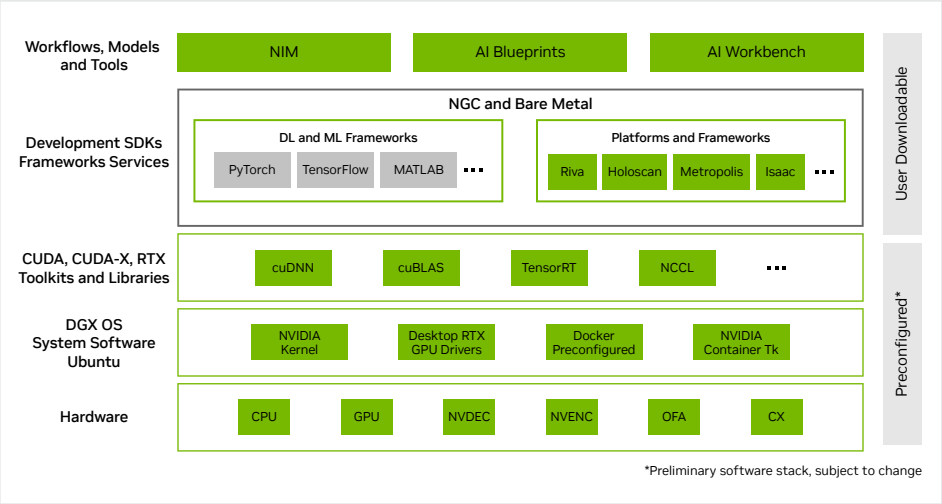
主要特性

- 基于 NVIDIA GB10 Grace Blackwell 超级芯片构建
- 配备 NVIDIA Blackwell GPU，支持第五代 Tensor Core 技术
- NVIDIA Grace CPU 采用高性能 20-core Arm 架构
- 支持 FP4，可提供高达 1 petaFLOP 的 AI 性能
- 128 GB 统一寻址系统内存
- 支持高达 200B 参数的模型
- NVIDIA ConnectX™ 网络可连接两台系统，从而支持对高达 405B 参数的模型进行处理
- 4 TB 的 NVMe 存储
- 小巧的桌面端外形

与新一代大参数 AI 模型协同工作

凭借 128 GB 的统一寻址系统内存和对 FP4 数据格式的支持，NVIDIA DGX Spark 可以支持多达 200B 参数的 AI 模型，使 AI 开发人员能够在桌面端对新一代 AI 推理模型进行原型设计、微调和推理。借助内置的 NVIDIA ConnectX 网络技术，可以连接两台 NVIDIA DGX Spark 系统，以处理 Llama 3.1 405B 等更大的模型。

本地开发，随时随地进行大规模部署



NVIDIA DGX Spark 软件堆栈

NVIDIA DGX Spark 为企业组织和开发者提供了一个功能强大且经济实惠的实验场地，用于原型设计模型，从而释放出更适合训练和部署生产模型的集群环境中宝贵的计算资源。NVIDIA AI 平台的软件架构支持 NVIDIA DGX Spark 用户将其工作任务从桌面端轻松迁移到 DGX Cloud 或任何加速云或数据中心基础设施，因此比以往都更容易进行原型设计、微调和迭代。

技术规格*	
架构	NVIDIA Grace Blackwell
GPU	NVIDIA Blackwell 架构
CPU	20 core Arm，10 Cortex-X925 + 10 Cortex-A725 Arm
CUDA 核心	NVIDIA Blackwell Generation
Tensor Core	第 5 代
RT Core 核心	第 4 代
Tensor Performance[1]	1 PFLOP
系统内存	128 GB LPDDR5x，一致性统一寻址系统内存
内存接口	256-bit
内存位宽	高达 273 GB/秒
存储	4 TB NVME.M2 with self-encryption
USB	4 个 USB Type C
以太网	1 个 RJ-45 接口 10 GbE
网卡	ConnectX-7 NIC @ 200 Gbps
Wi-Fi	WiFi 7
蓝牙	BT 5.4 w/LE
音频输出	HDMI 多通道音频输出
功耗	240 W
显示器接口	1 个 HDMI 2.1a
NVENC | NVDEC	1x | 1x
操作系统	NVIDIA DGX™ OS
尺寸	150 mm L x 150 mm W x 50.5 mm H
重量	1.2 kg

* 初步规格，可能会发生变化

准备好开始了吗?

如需详细了解 NVIDIA DGX Spark 系统，请访问 nvidia.cn/products/workstations/dgx-spark

1. 基于稀疏特性的理论 FP4 TOPS 算力
2. 使用 FP4 精度模型