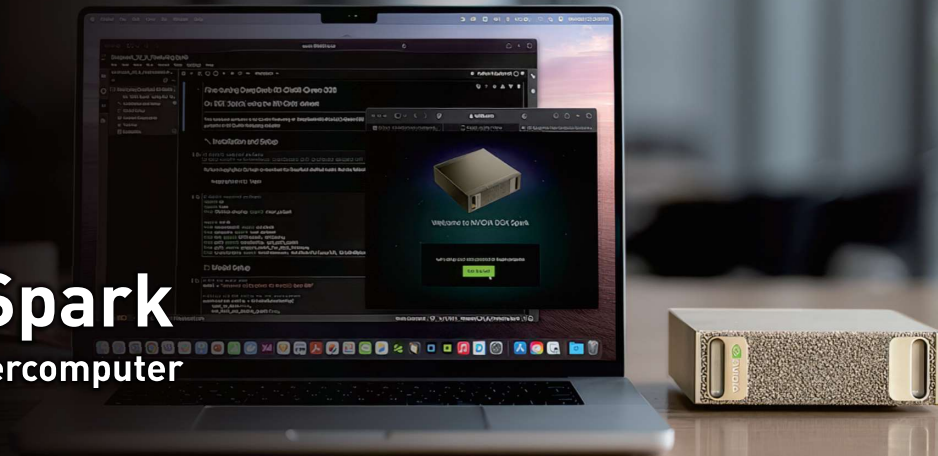




NVIDIA DGX Spark

A Grace Blackwell AI supercomputer on your desk.



デスクトップ AI コンピューティングの必要性

生成 AI モデルの規模と複雑さの増大により、ローカルシステムでの開発作業は困難を極めています。大規模モデルのローカルでのプロトタイピング、チューニング、推論には、大量のメモリと高度なコンピューティング性能が必要です。企業、ソフトウェアプロバイダー、政府機関、スタートアップ企業、そして研究者が AI 開発に注力するにつれ、AI コンピューティングリソースの必要性は高まり続けています。

デスク上の 200B パラメータモデル

NVIDIA DGX™ Spark は、AI の構築と実行のためにゼロから設計された新しいクラスのコンピューターです。NVIDIA GB10 Grace Blackwell スーパーチップを搭載し、NVIDIA Grace Blackwell アーキテクチャをベースとする NVIDIA DGX Spark は、最大 1000 TOPS の AI パフォーマンスを実現し、大規模な AI ワークロードを強力にサポートします。128 GB の統合システムメモリにより、開発者は最大 2000 億個のパラメータを持つモデルの実験、微調整、推論を行うことができます。さらに、NVIDIA ConnectX™ ネットワークにより、2 台の NVIDIA DGX Spark スーパーコンピューターを接続することで、最大 4050 億個のパラメータを持つモデルでの推論が可能になります。

開発者に使い慣れたエクスペリエンスを提供するために、NVIDIA DGX Spark は、産業用 AI ファクトリーを支えるのと同じソフトウェアアーキテクチャを採用しています。Ubuntu Linux で最新の NVIDIA AI ソフトウェアスタックが事前構成された NVIDIA DGX OS と、NVIDIA NIM™ および NVIDIA Blueprints への開発者プログラム アクセスを使用することで、開発者は Pytorch、Jupyter、Ollama などの一般的なツールを使用して NVIDIA DGX Spark 上でプロトタイプ作成、微調整、推論を行い、データセンターやクラウドにシームレスに展開できます。

NVIDIA DGX Spark はコンパクトなパッケージで圧倒的なパフォーマンスと機能を提供することで、開発者、研究者、データサイエンティスト、学生が生成 AI の限界を押し広げ続けることを可能にします。

NVIDIA Grace Blackwellをベースに構築

NVIDIA DGX Spark の中核を成すのは、デスクトップフォームファクター向けに最適化された NVIDIA Grace Blackwell アーキテクチャをベースにした新しい NVIDIA GB10 Grace Blackwell スーパーチップです。GB10 は、第 5 世代 Tensor コアと FP4 サポートを備えた強力な NVIDIA Blackwell GPU を搭載し、最大 1000 TOPS の AI コンピューティングを実現します。GB10 は、高性能な Grace 20 コア Arm CPU を搭載し、データの前処理とオーケストレーションを強化し、モデルチューニングとリアルタイム推論を高速化します。GB10 スーパーチップは、NVLink™-C2C を使用することで、PCIe Gen 5 の 5 倍の帯域幅を備えた CPU+GPU コヒーレントメモリ モデルを提供します。

主な特徴

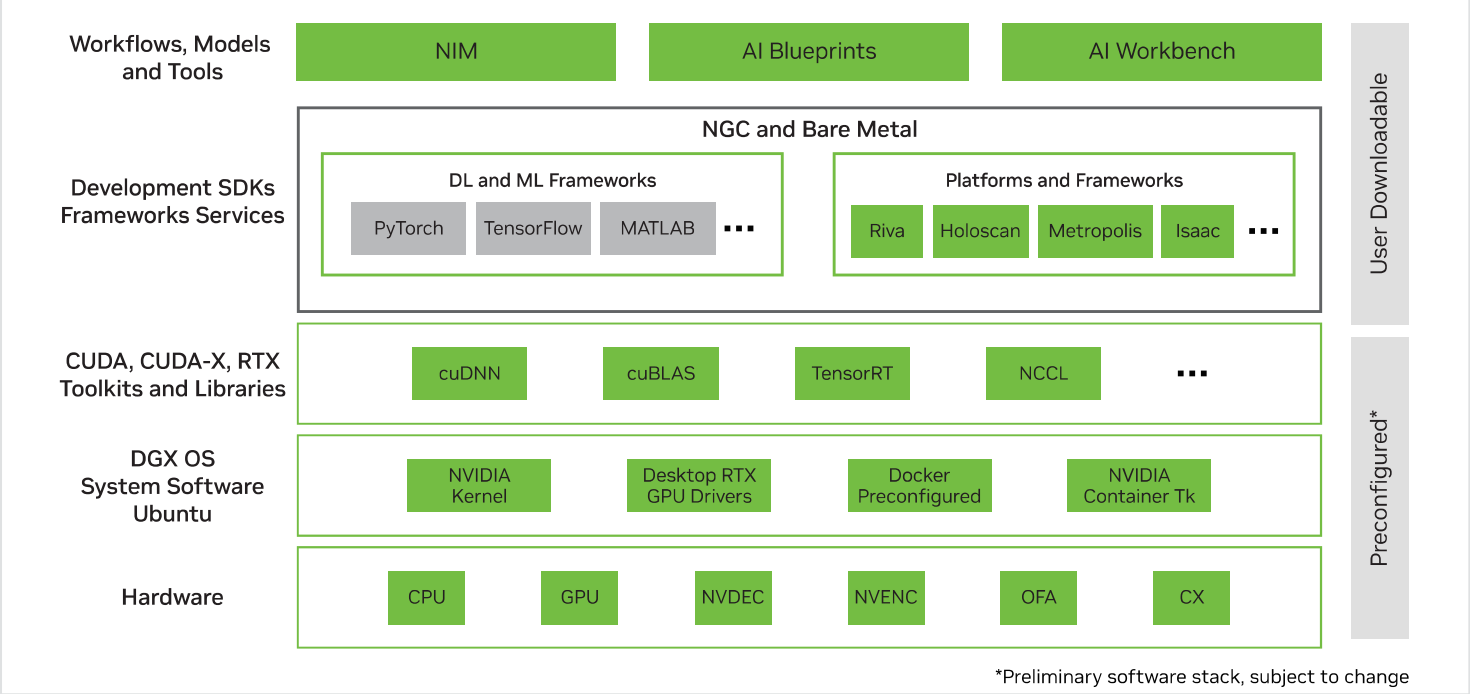
NVIDIA DGX Spark

- > NVIDIA GB10 Grace Blackwell スーパーチップ搭載
- > 第5世代 Tensor Core テクノロジー搭載 NVIDIA Blackwell GPU
- > 20コアの高性能 Arm アーキテクチャ搭載 NVIDIA Grace CPU
- > FP4 を使用した最大 1000 TOPS の AI パフォーマンス
- > 128 GB のコヒーレント統合システムメモリ
- > 最大 2,000 億個のパラメータモデルをサポート
- > 2 つのシステムをリンクし、最大 4,050 億個のパラメータモデルを処理できる NVIDIA ConnectX™ ネットワーク

大規模パラメータ AI モデルの操作

128GB の統合システムメモリと FP4 データフォーマットのサポートを備えた NVIDIA DGX Spark は、最大200Bのパラメータを持つ AI モデルをサポートし、AI開発者はデスクトップ上で大規模モデルのプロトタイプ作成、ファインチューニング、推論を行うことができます。NVIDIA ConnectX ネットワークテクノロジーを内蔵しているため、2台の NVIDIA DGX Spark システムを接続し、Llama 3.1 405B などさらに大規模なモデルで作業できます。

ローカルで開発し、大規模にどこにでも展開



NVIDIA DGX Spark ソフトウェアスタック

大規模パラメータ AI モデルの操作

NVIDIA DGX Spark は、組織や開発者にプロトタイプモデルのための強力かつ経済的な実験環境を提供し、クラスター環境の貴重なコンピューティングリソースを解放して、実稼働モデルのトレーニングと展開により適した環境を実現します。NVIDIA AI プラットフォームのソフトウェアアーキテクチャを活用することで、NVIDIA DGX Spark ユーザーは、コード変更をほとんど必要とせずに、デスクトップから DGX Cloud や高速クラウド、データセンターインフラストラクチャにモデルをシームレスに移行できるため、プロトタイプの実験、ファインチューニング、反復処理がこれまで以上に容易になります。

スペック (掲載情報は予告なく変更されることがございます)

Architecture	NVIDIA Grace Blackwell	USB	4x USB TypeC
GPU	NVIDIA Blackwell Architecture	Ethernet	1x RJ-45 connector 10 GbE
CPU	20 core Arm, 10 Cortex-X925 + 10 Cortex-A725 Arm	NIC	ConnectX-7 Smart NIC
CUDA Cores	NVIDIA Blackwell Generation	Wi-Fi	WiFi 7
Tensor Cores	5th Generation	Bluetooth	BT 5.3 w/LE
RT Cores	4th Generation	Audio-output	HDMI multichannel audio output
Tensor Performance ¹	1000 AI TOPS	Power Consumption	TBD
System Memory	128 GB LPDDR5x, unified system memory	Display Connectors	1x HDMI 2.1a
Memory Interface	256-bit	NVENC NVDEC	1x 1x
Memory Bandwidth	273 GB/s	OS	NVIDIA DGX™ OS
Storage	1 or 4 TB NVME.M2 with self-encryption	System Dimensions	150 mm L x 150 mm W x 50.5 mm H
		System Weight	1.2 kg

1. Theoretical FP4 TOPS using the sparsity feature.



株式会社 HPCテック

本社：〒103-0006 東京都中央区日本橋富沢町 7-13
TEL: 03-5643-2681 FAX: 03-5643-2682
大阪営業所：〒532-0011 大阪市淀川区西中島4丁目5-1
TEL 06-6195-6464 FAX 06-6195-6468

<https://www.hpctech.co.jp>
sales@hpctech.co.jp