Datasheet

NVIDIA DGX Spark

DGX パーソナル AI コンピューター、
AIの構築と実行を目的として設計

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。

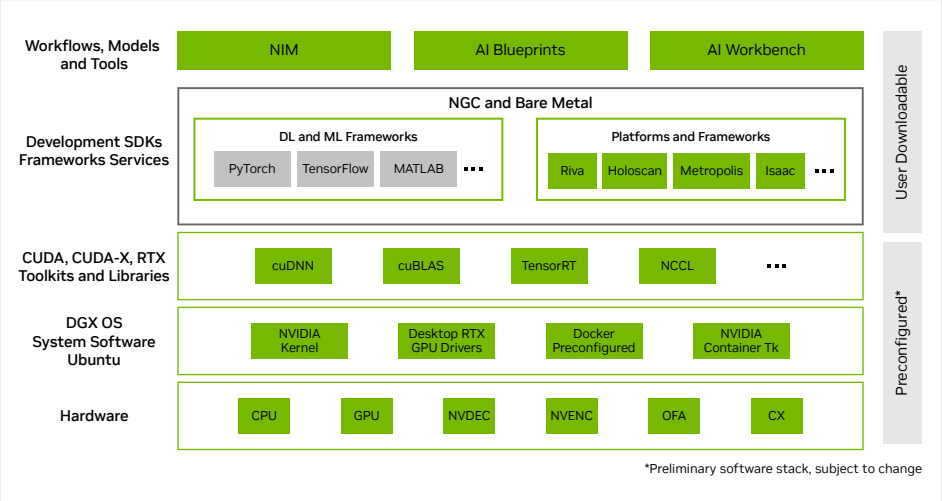
主な特徴

- NVIDIA 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
- 最大 4 TB の NVMe ストレージ
- コンパクトなデスクトップ フォームファクター

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 などの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

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、プロトタイプの作成、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。

技術仕様*

アーキテクチャ	NVIDIA Grace Blackwell
GPU	NVIDIA Blackwell アーキテクチャ
CPU	20 コア Arm, 10 Cortex-X925 + 10 Cortex-A725 Arm
CUDA コア	NVIDIA Blackwell 世代
Tensor コア	第5世代
RT コア	第4世代
Tensor パフォーマンス[1]	1000 AI TOPS
システムメモリ	128 GB LPDDR5x, 統合システムメモリ
メモリインタフェース	256-bit
メモリ帯域幅	273 GB/s
ストレージ	1 TB または 4 TB NVME.M2 自己暗号化機能付き
USB	4x USB TypeC
イーサネット	1x RJ-45 コネクタ 10 GbE
NIC	ConnectX-7 Smart NIC
Wi-Fi	WiFi 7
Bluetooth	BT 5.3 w/LE
音声出力	HDMI マルチチャンネルオーディオ出力
消費電力	TBD
ディスプレイコネクタ	1x HDMI 2.1a
NVENC | NVDEC	1x | 1x
OS	NVIDIA DGX™ OS
システム外形寸法	150 mm L x 150 mm W x 50.5 mm H
システム重量	1.2 kg

* 暫定仕様、変更の可能性があります。
1. スパース性機能を使用した理論的な FP4 TOPS.

始める準備はできましたか？

NVIDIA DGX Spark について更に詳しい情報は
www.nvidia.com/ja-jp/products/workstations/dgx-spark/

サーヴァンツ インターナショナル 株式会社
160-0023 東京都新宿区西新宿
コンシェリア西新宿タワーズウエスト 4 F
電話: 03-4455-7531 F
www.servants.co.jp