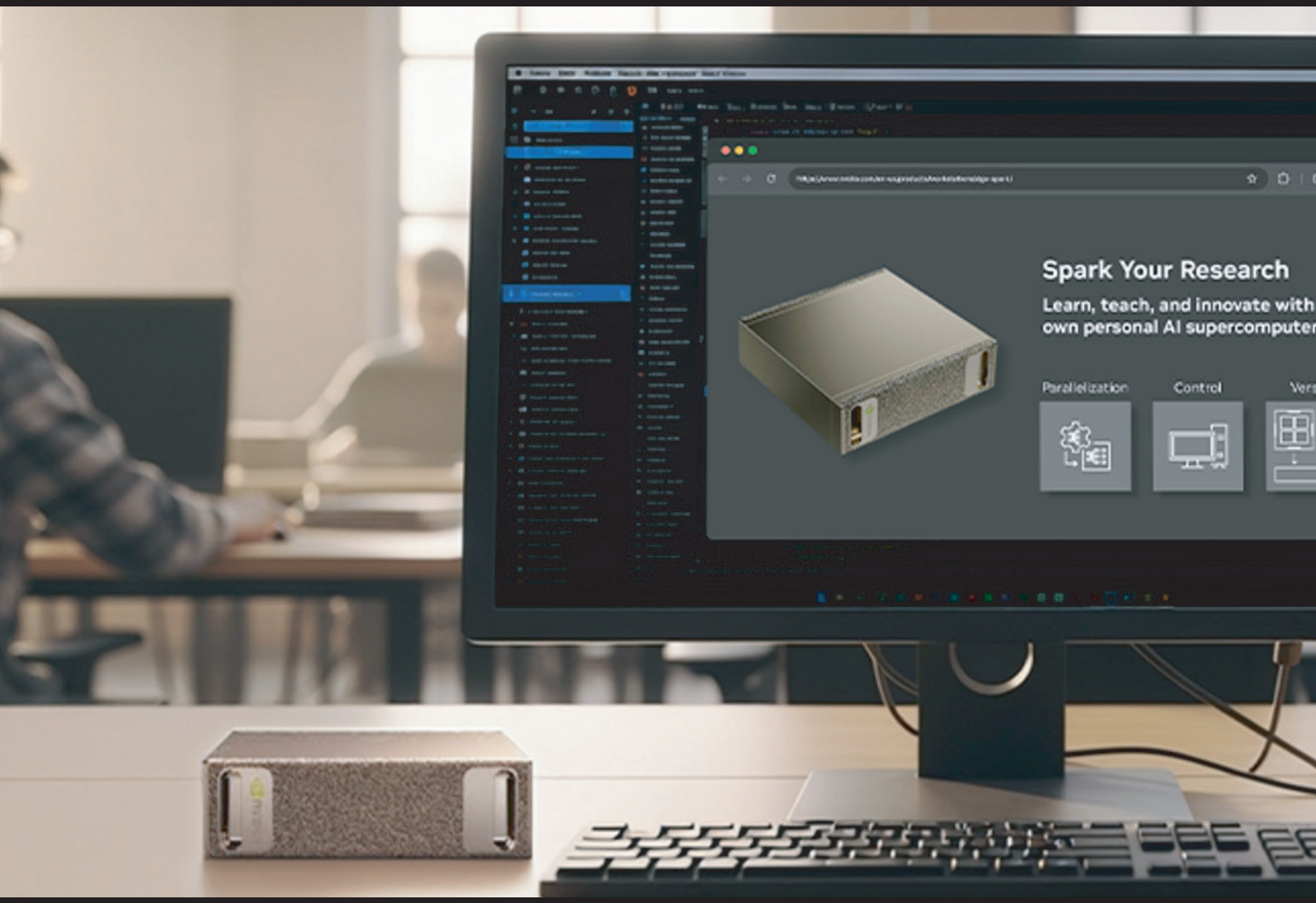
BN i&C | NVIDIA

내 책상 위의 Grace Blackwell AI 슈퍼컴퓨터

NVIDIA DGX SPARK

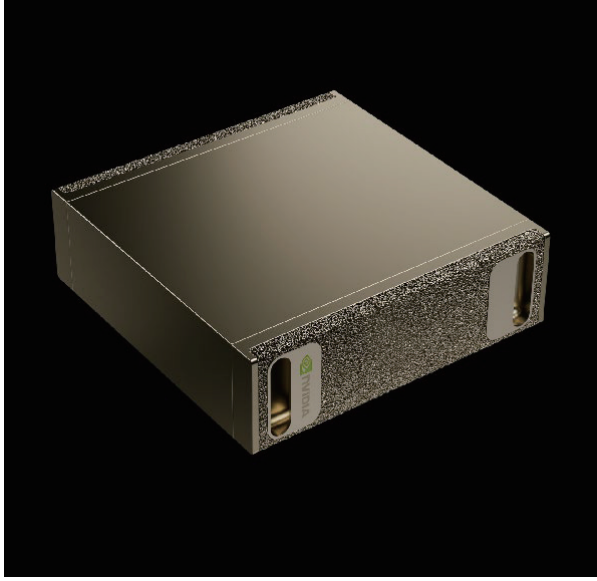


BN i&C

데이터시트

NVIDIA DGX SPARK

AI를 구축하고 실행하도록 설계된
DGX 개인용 AI 컴퓨터



데스크톱 AI 컴퓨팅 수요

생성형 AI 모델의 규모와 복잡성이 증가함에 따라, 로컬 시스템에서의 개발이 점점 더 어려워지고 있습니다. 대규모 모델을 로컬에서 프로토타이핑, 튜닝 및 추론하려면 많은 양의 메모리와 상당한 컴퓨팅 성능이 필요합니다. 기업, 소프트웨어 제공업체, 정부 기관, 스타트업, 연구원들이 AI 개발 인력을 확충함에 따라 AI 컴퓨팅 리소스에 대한 수요도 계속해서 증가하고 있습니다.

책상 위 2천억 개의 파라미터 모델

NVIDIA DGX™ Spark는 AI 구축 및 실행을 위해 처음부터 설계된 새로운 종류의 컴퓨터입니다. NVIDIA GB10 Grace Blackwell 슈퍼칩과 NVIDIA Grace Blackwell 아키텍처를 기반으로 하는 NVIDIA DGX Spark는 최대 1페타플롭의 AI 성능을 제공하여 대규모 AI 워크로드를 지원합니다. 개발자는 128GB의 통합 시스템 메모리를 통해 최대 2천억 개 파라미터의 모델을 실험, 파인튜닝 또는 추론할 수 있습니다. 또한 NVIDIA ConnectX™ 네트워킹을 통해 두 대의 NVIDIA DGX Spark 슈퍼컴퓨터를 연결하여 최대 4천5십억 개 파라미터의 모델까지 추론할 수 있습니다.

개발자가 친숙하게 사용할 수 있도록, NVIDIA DGX Spark는 산업용 AI 팩토리를 구동하는 동일한 소프트웨어 아키텍처를 제공합니다. NVIDIA DGX OS와 Ubuntu Linux를 사용하고, 최신 NVIDIA AI 소프트웨어 스택이 사전 구성되어 있으며, 개발자 프로그램을 통한 NVIDIA NIM™ 및 NVIDIA Blueprints 액세스가 가능하므로, 개발자는 PyTorch, Jupyter, Ollama와 같은 일반적인 도구를 사용하여 DGX Spark에서 프로토타이핑, 파인튜닝 및 추론하고, 이를 데이터센터나 클라우드에 원활하게 배포할 수 있습니다.

컴팩트한 패키지에 방대한 성능과 기능을 제공하는 NVIDIA DGX Spark를 통해 개발자, 연구원, 데이터 사이언티스트, 학생은 생성형 AI의 경계를 계속해서 넓혀 나갈 수 있습니다.

NVIDIA Grace Blackwell 기반

NVIDIA DGX Spark의 핵심에는 데스크톱 폼팩터에 최적화된 NVIDIA Grace Blackwell 아키텍처 기반의 새로운 NVIDIA GB10 Grace Blackwell 슈퍼칩이 있습니다. GB10은 5세대 Tensor 코어와 FP4를 지원하는 강력한 NVIDIA Blackwell GPU를 탑재하여 최대 1,000TOPS의 AI 컴퓨팅 성능을 제공합니다. GB10에는 데이터 전처리 및 오케스트레이션을 강화하여 모델 튜닝 및 실시간 추론 속도를 높이는 고성능 Grace 20코어 Arm CPU가 탑재되어 있습니다. GB10 슈퍼칩은 NVLink™-C2C를 사용하여 PCIe Gen 5 대비 5배 대역폭을 지원하는 일관된 CPU+GPU 메모리 모델을 제공합니다.

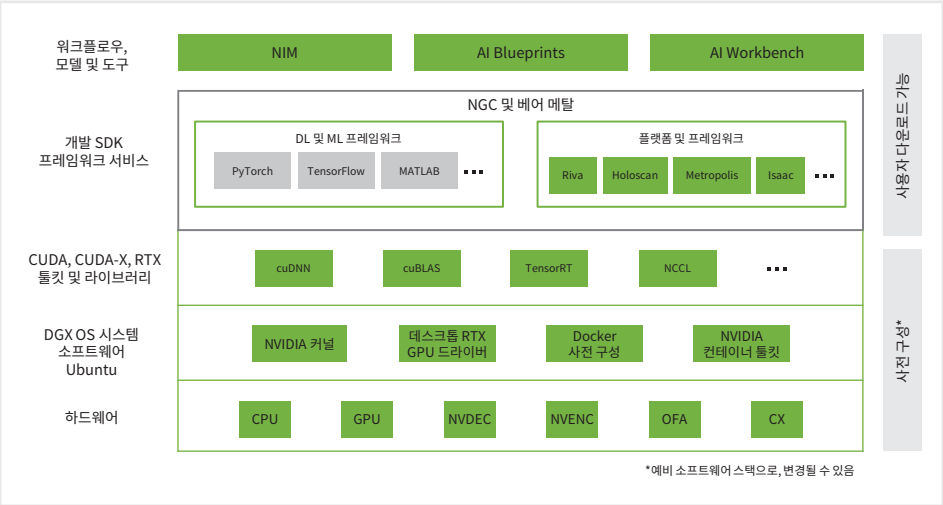
주요 특징

- NVIDIA GB10 Grace Blackwell 슈퍼칩 기반
- 5세대 Tensor 코어 기술 지원 NVIDIA Blackwell GPU
- 20코어 고성능 Arm 아키텍처의 NVIDIA Grace CPU
- FP4를 사용한 최대 1페타플롭의 AI 성능
- 128GB의 일관된 통합 시스템 메모리
- 최대 2천억 개 파라미터 모델 지원
- 두 시스템을 연결하여 최대 4천5십억 개 파라미터 모델까지 처리 가능한 NVIDIA ConnectX™ 네트워킹
- 최대 4TB의 NVMe 스토리지
- 컴팩트한 데스크톱 폼 팩터

대규모 파라미터 AI 모델 작업

128GB의 통합 시스템 메모리와 FP4 데이터 형식 지원을 통해 NVIDIA DGX Spark는 최대 2천억 개 파라미터의 AI 모델을 지원 가능하므로, AI 개발자가 데스크톱에서 대규모 모델을 프로토타이핑, 파인튜닝 및 추론할 수 있습니다. 내장된 NVIDIA ConnectX 네트워크 기술을 통해 두 대의 NVIDIA DGX Spark 시스템을 연결하여 Llama 3.1 405B와 같은 훨씬 더 대규모의 모델 작업도 가능합니다.

로컬에서 개발, 어디서나 대규모 배포



NVIDIA DGX Spark 소프트웨어 스택

NVIDIA DGX Spark는 조직과 개발자에게 프로토타입 모델을 위한 강력하고 경제적인 실험 환경을 제공하여 클러스터 환경에서 프로덕션 모델을 훈련하고 배포하는 데 더 적합한 귀중한 컴퓨팅 리소스를 확보할 수 있게 해줍니다. NVIDIA AI 플랫폼 소프트웨어 아키텍처를 활용하면 NVIDIA DGX Spark 사용자는 거의 코드 변경 없이 데스크톱에서 DGX 클라우드 또는 가속화된 클라우드 또는 데이터센터 인프라로 모델을 원활하게 이동할 수 있으므로, 프로토타이핑, 파인튜닝 및 반복 작업을 그 어느 때보다 쉽게 수행할 수 있습니다.

기술 사양*

아키텍처	NVIDIA Grace Blackwell
GPU	NVIDIA Blackwell 아키텍처
CPU	20코어 Arm, 10 Cortex-X925 + 10 Cortex-A725 Arm
CUDA 코어	NVIDIA Blackwell 세대
Tensor 코어	5세대
RT 코어	4세대
Tensor 성능[1]	1페타플롭
시스템 메모리	128GB LPDDR5x, 통합 시스템 메모리
메모리 인터페이스	256비트
메모리 대역폭	273GB/s
스토리지	1TB 또는 4TB NVMe M.2(자체 암호화 지원)
USB	4개의 USB TypeC
이더넷	1개의 RJ-45 커넥터 10GbE
NIC	ConnectX-7 Smart NIC
Wi-Fi	WiFi 7
블루투스	BT 5.3(저전력 지원)
오디오 출력	HDMI 멀티채널 오디오 출력
전력 소모량	미정
디스플레이 커넥터	1개의 HDMI 2.1a
NVENC | NVDEC	1개 | 1개
OS	NVIDIA DGX™OS
시스템 크기	150mm(길이) x 150mm(너비) x 50.5mm(높이)
시스템 무게	1.2kg

* 예비 사양으로, 변경될 수 있음

1. 희소성 기능을 사용한 이론적 FP4 TOPS입니다.

Ready to Get Started?

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

BN I&C | NVIDIA