



# NVIDIA DGX GB300

최첨단 AI 모델 가속화를 위한 성능 한계의 돌파구



## AI 추론 시대에 맞는 목적 기반의 AI 팩토리 인프라

최근 몇 년, 생성형 AI와 거대 언어 모델(LLM)은 폭발적인 성장을 이루며 AI 지형을 근본적으로 바꾸어 놓았습니다. 이에 기업들은 전사적 AI 전략을 빠르게 수립하고 있습니다. 이러한 기하급수적 확장과 함께 새로운 과제들도 등장하기 시작했습니다. 특히, 프런티어 모델에서 요구되는 고성능 연산이 가능한 인프라의 중요성이 더욱 커졌습니다. 최근의 AI 모델은 방대한 크기와 복잡성으로 인해 고성능 분산 컴퓨팅 클러스터와 고급 네트워킹, 대규모 메모리 풀을 필요로 합니다. AI 지형이 계속해서 진화함에 따라, 기업이 AI 혁신의 선두를 유지하고, 산업 전반에 걸쳐 기술 돌파구를 이끌어, 고급 AI 애플리케이션을 통해 비즈니스 전환을 할 수 있는 새로운 가능성들을 열기 위해서는 목적 기반 인프라 솔루션의 역할이 핵심적입니다.

NVIDIA DGX™ GB300은 이러한 AI 추론 시대를 위해 설계된 종합적 AI 인프라 솔루션으로, 최첨단 모델의 훈련과 추론에 최적화되어 있습니다. Grace Blackwell Ultra 슈퍼칩 기반으로 구축되어, NVIDIA DGX SuperPOD와 함께 수만 개의 슈퍼칩으로 확장 가능함에 따라, 방대한 공유 메모리 공간이 형성 가능하여 세계 최대 규모의 AI 모델도 가속화가 가능합니다. 또한 랙 스케일의 100% 수냉식 설계는 NVIDIA MGX 랙을 사용하는 현대의 데이터 센터에 맞게 최적화되어, 기업의 고성능 AI 하드웨어를 관리하는 부담을 줄여줍니다. DGX GB300은 기업이 성능 한계를 돌파하고 가장 까다로운 AI 워크로드를 처리할 수 있도록 지원하는 인프라 솔루션입니다.

## NVIDIA Grace Blackwell Ultra 기반의 설계

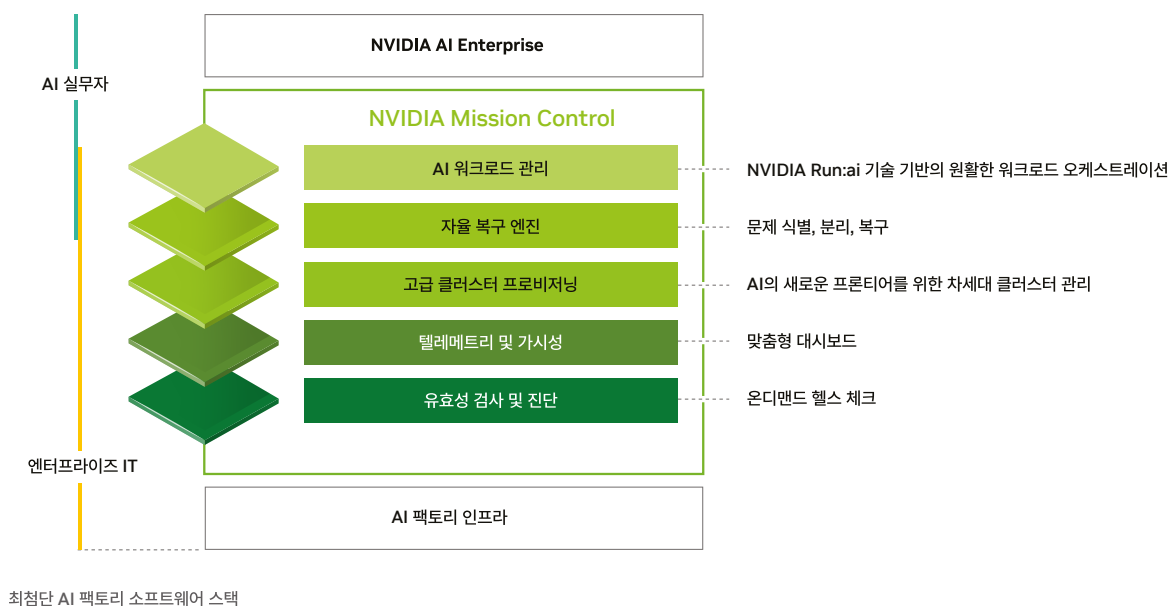
DGX GB300은 NVIDIA Grace Blackwell Ultra 아키텍처에 기반하여 설계되어, AI 컴퓨팅 성능에서 획기적인 도약을 이루었습니다. 전례 없는 추론 성능과 함께, 어떤 AI 워크로드에서도 향상된 훈련 성능을 제공하는 최첨단 시스템입니다. Grace CPU 36개와 Blackwell Ultra GPU 72개가 NVIDIA NVLink 스위치 시스템을 통해 하나의 거대한 GPU처럼 연결되어, 1.4 exaFLOPS의 추론 성능과 360 petaFLOPS의 훈련 성능을 제공합니다. DGX GB300은 모델 훈련, 훈련 후 최적화, 추론 테스트 단계까지 전체 AI 파이프라인의 성능을 최적화하도록 목적 기반으로 설계되어, 기업이 AI 추론 시대의 요구에 맞춰 인프라를 확장할 수 있도록 돕습니다.

### 주요 특징

- > NVIDIA GB300 Grace Blackwell Ultra 슈퍼칩 기반
- > NVIDIA DGX SuperPOD를 통해 수만 개의 GB300 슈퍼칩으로 확장 가능
- > 72개의 NVIDIA Blackwell Ultra GPU를 NVIDIA® NVLink®로 하나로 연결
- > 효율적인 100% 수냉식 랙 스케일 설계
- > NVIDIA 네트워킹
- > NVIDIA AI Enterprise 및 NVIDIA Mission Control 소프트웨어 활용

# NVIDIA Mission Control로 모델 실행과 핵심 업무 자동화

NVIDIA Mission Control은 세계 최고 수준의 운영 역량을 소프트웨어 형태로 조직에 제공하여, 개발자 워크로드부터 인프라, 시설 관리까지 AI 팩토리 운영의 모든 측면을 제어하는 플랫폼입니다. 이를 통해 훈련과 추론의 즉각적인 민첩성을 높이는 동시에, 인프라 복원력을 위해 풀스택 인텔리전스를 제공합니다. NVIDIA Mission Control은 어떤 기업이든 하이퍼스케일 수준의 효율성으로 AI를 구동하여 AI 실험을 가속화할 수 있도록 합니다. 또한, AI 개발과 배포를 간소화하는 소프트웨어 제품군인 NVIDIA AI Enterprise가 NVIDIA DGX 시스템에 최적화되어 있으며, NVIDIA NIM™ 마이크로서비스를 사용하여 속도, 사용 편의성, 관리 용이성, 보안성과 함께 최적의 모델을 배포할 수 있습니다.



## 현대의 데이터 센터를 위한 설계

DGX GB300은 현대의 데이터 센터 환경에 끊임 없이 통합되도록 설계되어, 탁월한 유연성과 확장성을 제공합니다. DGX GB300을 통해 기업은 기존 인프라를 전면 교체하지 않고도 하이퍼스케일러 수준의 운영이 가능합니다. 100% 수냉식 랙 설계를 통해 에너지 효율성을 획기적으로 향상시키며, 고급 AI 워크로드에 필수적인 고전력 밀도 구현을 가능하게 합니다. DGX GB300은 하이퍼스케일 수준의 성능과 기존 인프라와 호환성까지 갖춰, 기업에서 성능, 효율성, 호환성 중 어느 하나도 타협하지 않고 AI의 잠재력을 최대한 활용할 수 있도록 혁신적인 솔루션을 제공합니다.

Technical Specifications

DGX GB300	
GPU	72x NVIDIA Blackwell Ultra GPUs in Grace Blackwell Ultra Superchips
CPU Cores	2,592
GPU Memory HBM3e	20.1TB
Total Fast Memory	37.9TB
Performance	1,400 petaFLOPS of FP4 AI performance*
	700 petaFLOPS of FP8 AI performance*
	360 petaFLOPS of FP16 AI performance*
Networking	72x OSFP single-port NVIDIA ConnectX®-8 VPI with 800Gb/s NVIDIA InfiniBand
	18x dual-port NVIDIA BlueField®-3 VPI with 200Gb/s InfiniBand and Ethernet
NVIDIA NVLink Switch System	9x L1 NVIDIA NVLink Switches
Management Network	Host baseboard management controller (BMC) with RJ45
Software	NVIDIA Mission Control
	NVIDIA AI Enterprise
	NVIDIA DGX OS
	Supports Ubuntu
Support	Three-year business-standard hardware and software support

\*Shown with sparsity.

Ready to Get Started?

To learn more about DGX GB300, visit [nvidia.com/dgx-gb300](https://nvidia.com/dgx-gb300)



© 2025 NVIDIA Corporation and affiliates. All rights reserved. NVIDIA, the NVIDIA logo, BlueField, ConnectX, DGX, DGX BasePOD, DGX SuperPOD, Grace, Mission Control, NIM, and NVLink 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. 3725800. MAR25