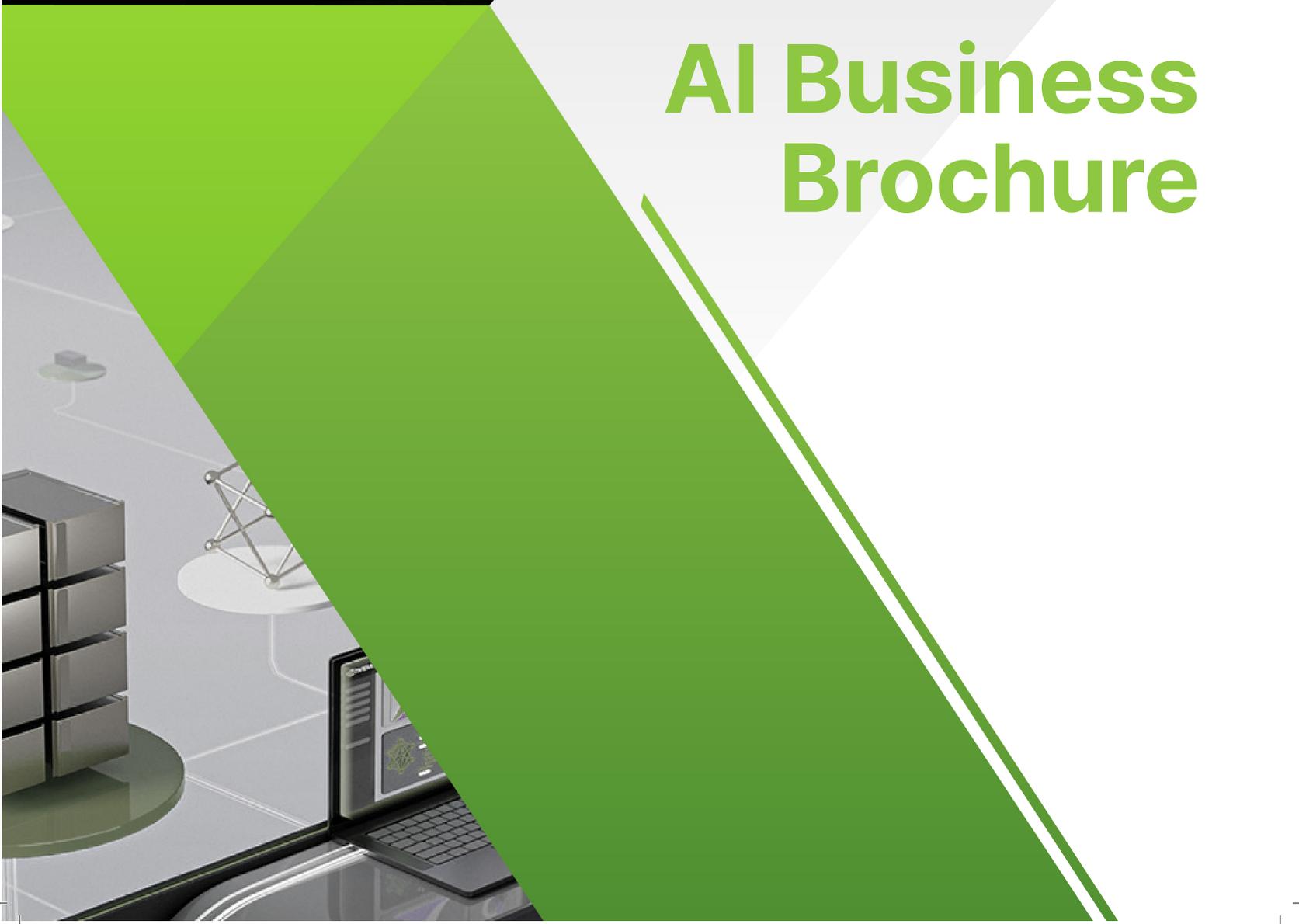




ICRAFT
Do it AI



AI Business Brochure



Contents

01 iCRAFT AI SERVICE

- AI Service & Data Platform Technologies 1
- AI Service Process 2

02 AI Software

- NVIDIA AI Enterprise 3
- NVIDIA Omniverse Enterprise 3

03 AI Hardware

- NVIDIA DGX SUPERPOD 4
- NVIDIA DGX B200 Server 5
- NVIDIA DGX H200 Server 6
- NVIDIA GPU 7
- NVIDIA Network 8
- NVIDIA DPU 12

04 Business Partner & Customer Base 13

iCRAFT AI SERVICE

AI Service & Data Platform Technologies

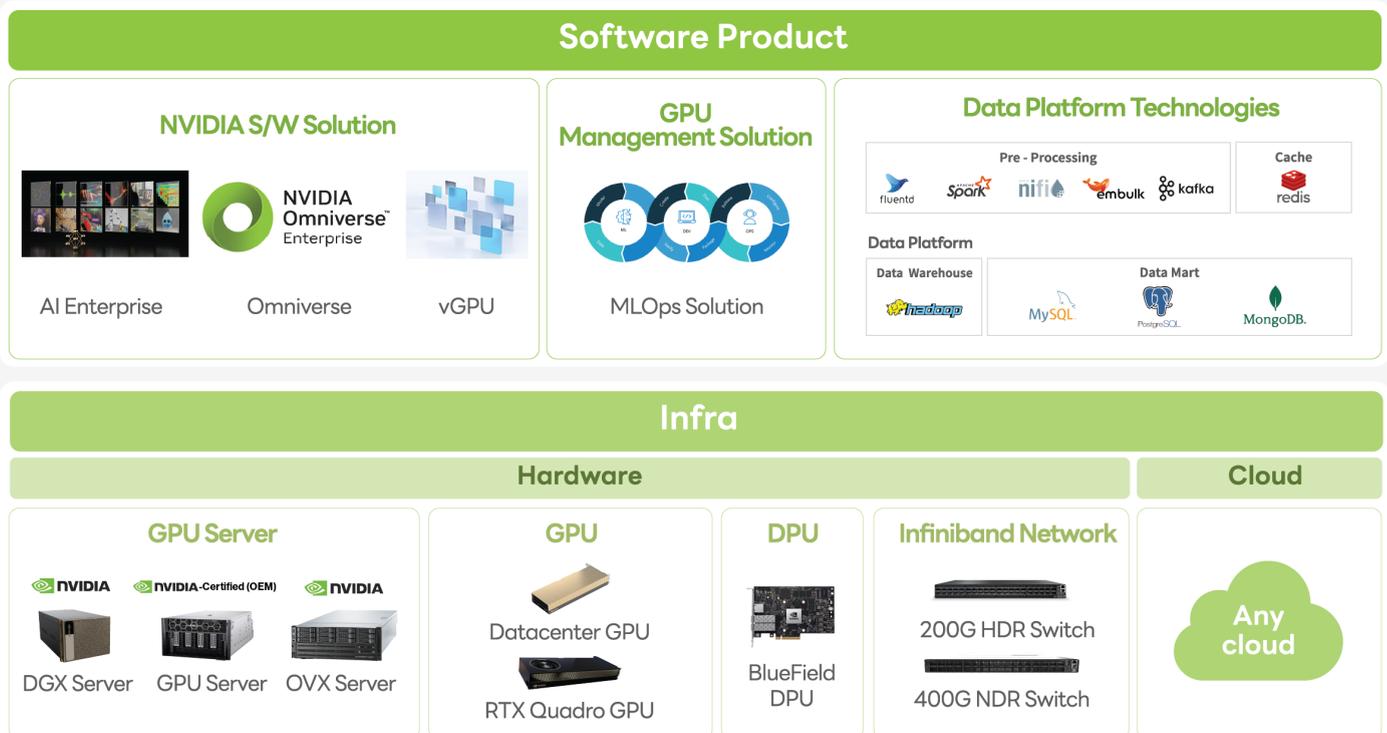


데이터 전처리과정부터 AI 시스템 구축까지.
"Full-Stack AI 공급사" 아이크래프트에 맡기세요!

아이크래프트는 NVIDIA 의 Elite 파트너이자 Full-Stack AI 공급사로서, 기업의 AI 서비스에 필요한 <데이터 전처리과정부터 AI 시스템구축까지> 전 단계 지원은 물론, 시스템구축 이후 운영에 대한 기술지원까지 제공합니다. Full-Stack AI 의 범위에는 하드웨어, 소프트웨어, 클라우드 및 플랫폼까지 포함되어 있습니다.

앞으로의 AI 프로젝트는 더욱더 특정 조직이나 산업의 고유한 요구사항과 과제에 맞는 맞춤형 AI 솔루션을 구축을 필요로 합니다. 이에 아이크래프트는 AI 선도기업인 NVIDIA 를 포함한 다양한 제조사 및 협력사와의 파트너십을 통해 기업이 AI 프로젝트를 통한 비즈니스 가치를 창출 할 수 있도록 데이터 수집단계부터 AI 시스템구축까지 One-stop 서비스를 지원합니다.

AI 시스템 구축을 위한 AI Full-stack 서비스 포트폴리오



AI SERVICE PROCESS

AI 서비스 컨설팅 및 제공 절차

iCRAFT FULL-STACK AI

아이크래프트는 FULL-STACK AI 서비스를 제공합니다.



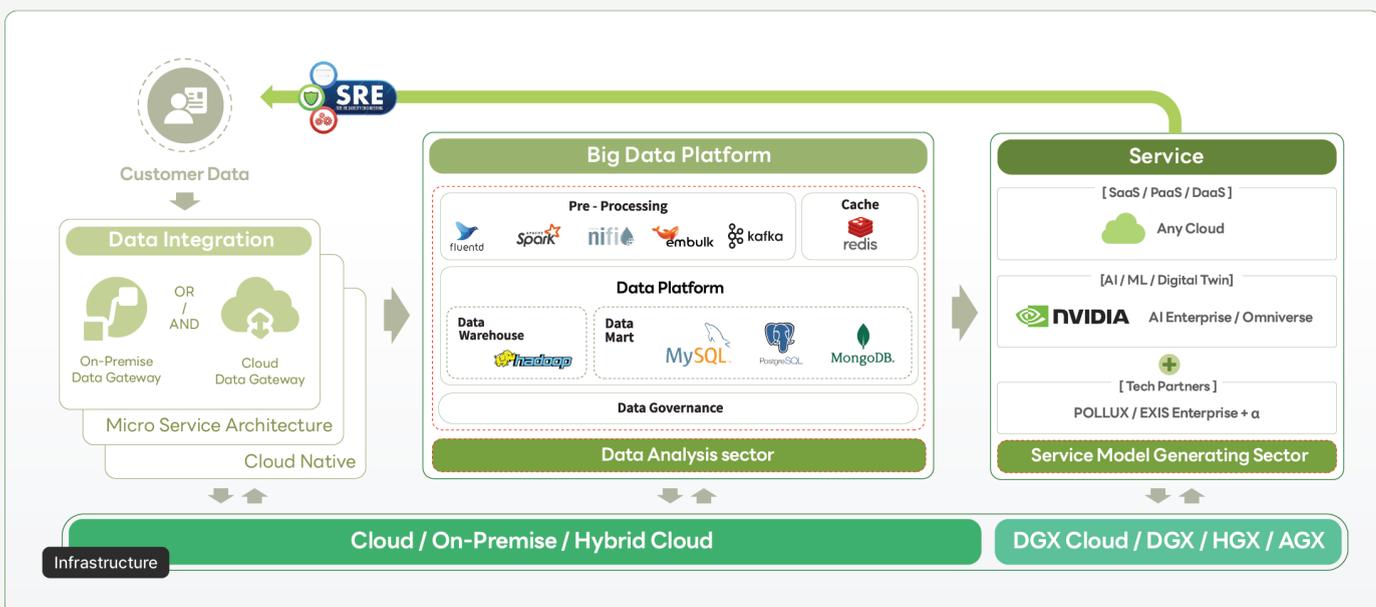
iCRAFT FULL-STACK AI Service 도입 기대효과

높은 성능

대규모 데이터 처리능력

서비스의 독립성 / 확장성

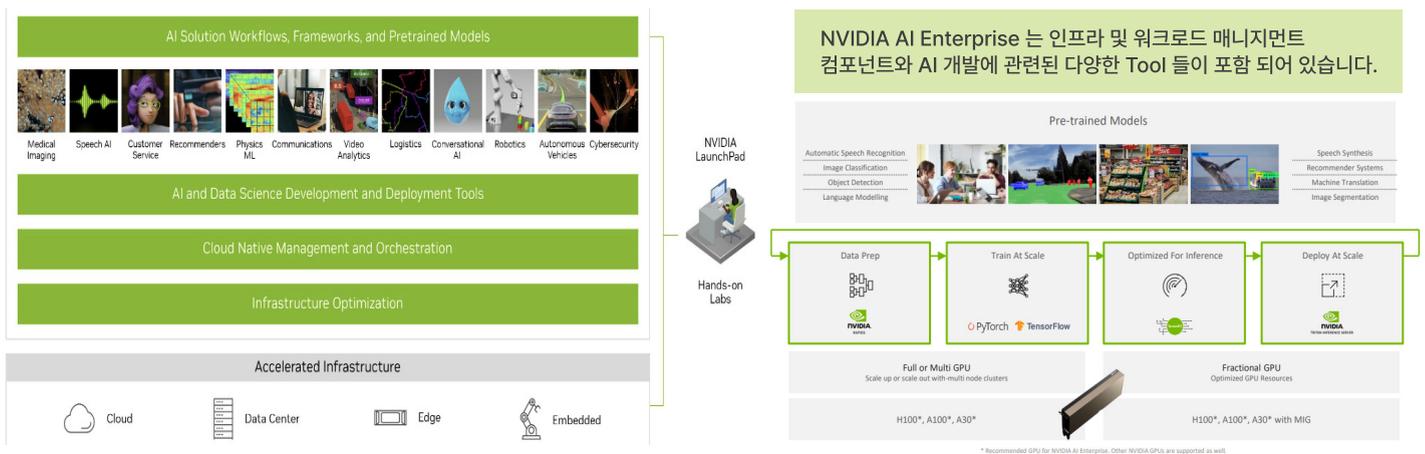
서비스의 안정성



NVIDIA AI Enterprise

프로덕션 AI용 End-to-End 소프트웨어 플랫폼

NVIDIA AI Enterprise는 데이터 처리를 가속화하고 생성 AI, 컴퓨터 비전, 음성 AI를 포함한 프로덕션 AI의 개발 및 배포를 간소화하기 위한 100개 이상의 프레임워크, 프리트레인드 모델 및 개발 도구를 제공합니다. 또한, NVIDIA AI Enterprise support 서비스를 통해 NVIDIA 엔터프라이즈 포트폴리오 전반에 걸쳐 전문 지식을 갖춘 숙련된 솔루션 엔지니어의 기술 지원을 제공합니다.



NVIDIA Omniverse Enterprise

메타버스 애플리케이션 제작 및 운영을 위한 플랫폼

NVIDIA Omniverse는 Universal Scene Description 기반 3D 워크플로우 및 애플리케이션을 개발할 수 있는 컴퓨팅 & 디지털트윈 플랫폼입니다.

01 빠른 협업을 위한 개방형 플랫폼

빠른 협업을 위해 탄생한 개방형 플랫폼인 NVIDIA Omniverse Enterprise는 확장 가능한 엔드 투 엔드 플랫폼으로 구축 및 운영을 지원합니다

USD(Universal Scene Description) 기반의 Omniverse Enterprise는 전 세계에 분산되어 있는 디자이너들의 협업을 가능하게 합니다.

02 NVIDIA Omniverse Enterprise 아키텍처



NVIDIA DGX SUPERPOD

세계 최초의 턴키 AI 데이터센터 솔루션



NVIDIA DGX SuperPOD™는 AI 데이터센터 인프라 플랫폼입니다. DGX SuperPOD는 업계에서 입증된 성능과 함께 AI 및 고성능 컴퓨팅(HPC) 워크로드를 위한 선도적인 표준 인프라 아키텍처와 확장성을 제공합니다.

엔터프라이즈 AI의 규모에 맞게 설계

NVIDIA DGX SuperPOD는 조직을 위한 턴키 AI 데이터센터 솔루션으로 세계적인 수준의 컴퓨팅, 소프트웨어 도구, 전문 지식, 지속적인 혁신을 원활하게 제공합니다. 두 가지 아키텍처 옵션을 갖춘 DGX SuperPOD를 통해 모든 엔터프라이즈는 플랫폼 복잡성으로 인해 어려움을 겪지 않으면서 AI를 비즈니스에 통합하고 혁신적인 애플리케이션을 만들 수 있습니다.

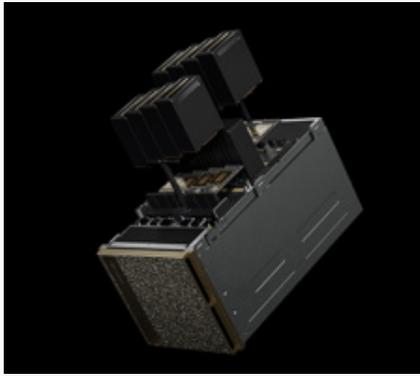


NVIDIA DGX B200 시스템을 활용하는 DGX SuperPOD

NVIDIA B200 시스템을 활용하는 DGX SuperPOD는 거대 언어 모델 구축, 공급망 최적화, 또는 대량의 데이터에서 인텔리전스 추출 등 복잡하고 다양한 AI 워크로드를 통해 모든 규모의 엔터프라이즈 팀을 지원하는 대규모 인프라에 이상적입니다.

NVIDIA DGX H200 시스템을 갖춘 DGX SuperPOD

NVIDIA DGX H200 시스템을 사용하는 DGX SuperPOD는 NVIDIA NeMo 프레임워크 및 딥 러닝 추천 시스템을 사용하는 거대 언어 모델(LLM)과 같이 크고 복잡하거나 트랜스포머 기반인 AI 워크로드를 지원하는 대규모 인프라에 가장 적합합니다.



NVIDIA DGX B200

AI Center of Excellence의 기반

NVIDIA DGX B200은 AI를 최대한 활용하기 위한 고성능의 컴퓨팅, 스토리지 및 네트워킹 능력을 충족하기 위해 출시된 NVIDIA DGX 플랫폼의 최신 제품입니다. 이 통합 AI 플랫폼은 NVIDIA Blackwell GPU와 고속 인터커넥트를 완벽하게 활용하여 AI 훈련, 파인 튜닝 및 추론을 위한 차세대 AI 성능을 제공합니다.

01 NVIDIA Blackwell GPU 탑재

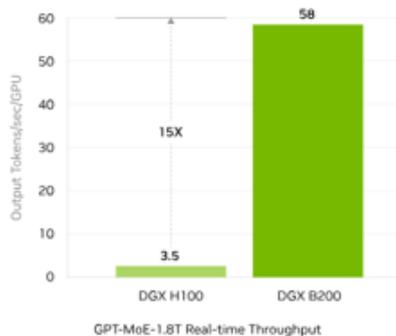
AI 확장의 장벽 돌파



NVIDIA DGX B200은 NVIDIA Blackwell GPU를 탑재한 최초의 시스템으로, 대규모 언어 모델과 자연어 처리와 같은 세계에서 가장 복잡한 AI 문제에 대해 획기적인 성능을 제공합니다. 8개의 Blackwell GPU로 구성된 DGX B200은 1.4TB의 GPU 메모리, 64TB/s의 HBM3e 메모리 대역폭, 그리고 14.4TB/s의 GPU 간 대역폭을 제공하여 AI 워크로드를 처리하는 데 이상적인 솔루션입니다.

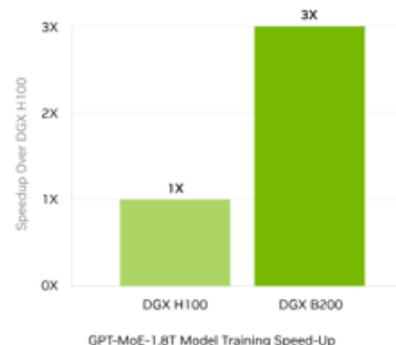
DGX B200 기반의 차세대 성능

실시간 거대 언어 모델 추론



토큰 간 지연 시간(TTL) = 실시간 50ms, 첫 토큰 지연 시간(FTL) = 5초, 입력 시퀀스 길이 = 32,768, 출력 시퀀스 길이 = 1,028, 8x 8방향 DGX H100 공병식 대 1x 8방향 DGX B200 공병식, GPU당 성능 비교입니다.

초고속 AI 학습 성능



32,768개의 GPU 규모, 4,096배 8방향 DGX H100 공병식 클러스터: 400G IB 네트워크, 4,096x 8방향 DGX B200 공병식 클러스터: 400G IB 네트워크.

* 예상 성능은 변경될 수 있습니다.

02 워크플로우 가속화를 위한 단일 플랫폼

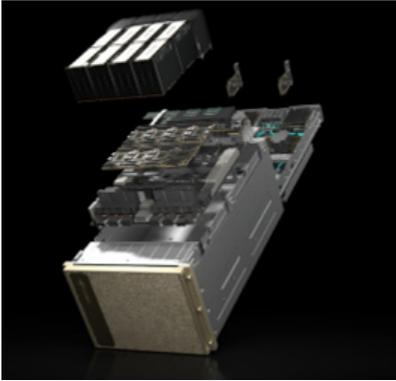
AI 워크플로우가 복잡해짐에 따라, 기업은 AI 파이프라인의 모든 단계에서 대규모 데이터셋을 처리할 수 있는 능력을 필요로 하게 되었습니다. NVIDIA DGX B200은 이러한 요구를 충족시키기 위해 설계되었으며, AI 워크플로우를 가속화하는 통합 플랫폼을 제공합니다.

03 최적화된 AI 소프트웨어 플랫폼 제공

엔터프라이즈 수준의 orchestration 기능을 제공하여 AI 개발 및 배포를 간소화 할 수 있는 NVIDIA Base Command 를 지원합니다.

NVIDIA DGX H200

세계적으로 입증된 엔터프라이즈 AI

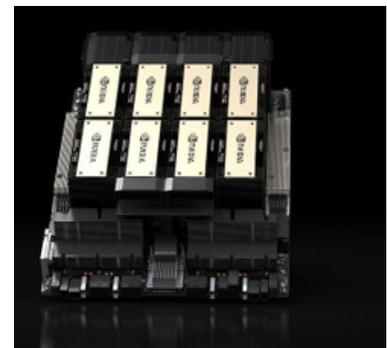


AI는 과학과 비즈니스 간의 격차를 해소했습니다. 이제 AI는 실험의 영역을 넘어 혁신과 비즈니스 최적화를 위해 매일 사용되고 있습니다. DGX H200은 세계 최초의 AI 전용 인프라 포트폴리오의 일환으로 설계되었으며, 기업 AI 센터의 중심이 되도록 설계되었습니다. 완전히 최적화된 하드웨어 및 소프트웨어 플랫폼이며, NVIDIA AI 소프트웨어 솔루션, 써드파티 및 NVIDIA 전문가에 대한 지원을 통해 가장 복잡한 비즈니스 문제를 해결할 수 있도록 합니다.

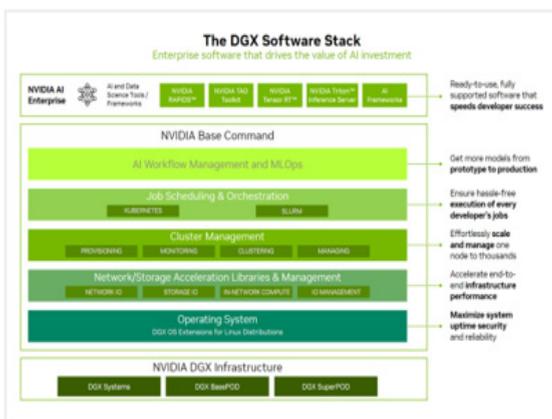
01 NVIDIA H200 TENSOR CORE GPU 탑재

AI 확장의 장벽 돌파

NVIDIA DGX H200은 AI 규모와 성능의 한계를 돌파합니다. 32 페타플롭스의 AI 성능과 NVIDIA ConnectX®-7 스마트 네트워크 인터페이스 카드(SmartNIC)를 통해 DGX A100보다 2배 빠른 네트워킹 성능을 제공합니다. 또한, NVIDIA DGX SuperPOD 및 DGX BasePOD을 위한 고속 확장성을 제공합니다. DGX H200은 가장 복잡한 AI 트레이닝 및 추론 작업을 위한 1,128GB의 GPU 메모리를 탑재하고 있습니다.



02 최적화된 AI 소프트웨어 플랫폼 제공



NVIDIA Base Command는 DGX 플랫폼을 지원하며, 기업이 NVIDIA 소프트웨어 혁신의 혜택을 누릴 수 있도록 합니다. 이 플랫폼은 엔터프라이즈 수준의 orchestration and cluster management 기능을 제공하여 AI 개발 및 배포를 간소화합니다.

NVIDIA GPU

GPU 카드 라인업

DataCenter GPU

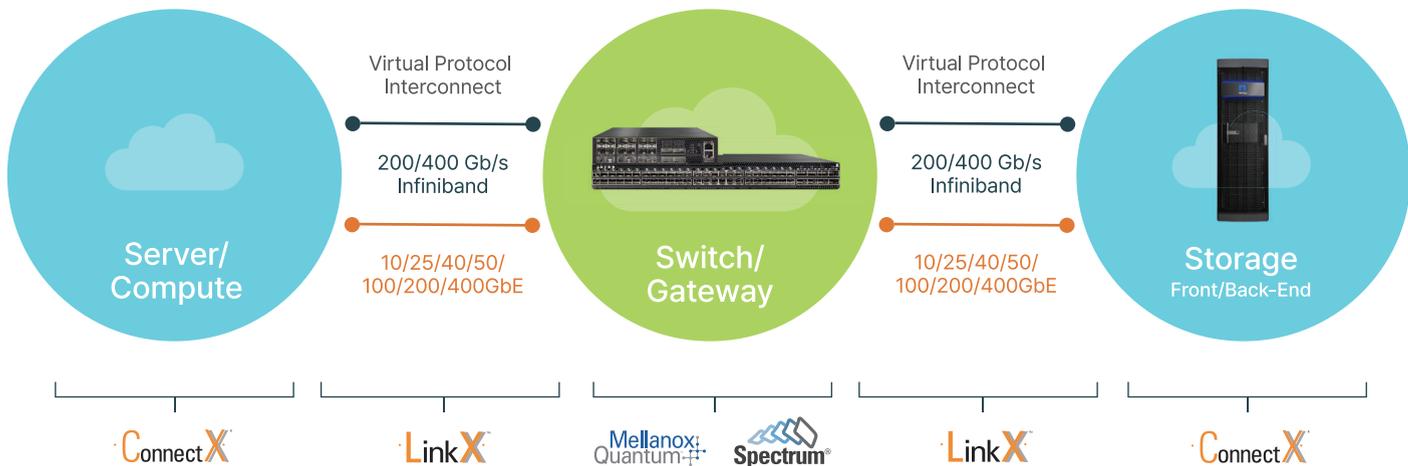
제품	H200 NVL	H100 NVL	L40S	L40	A10	A16
열처리	Passive					
메모리	141GB HBM3 ECC 지원	94GB HBM3 ECC 지원	48GB GDDR6 ECC 지원	24GB GDDR6 ECC 지원	24GB HBM2 ECC 지원	16GB GDDR6 ECC 지원
대역폭	4.8TB/s	3.9TB/s	864GB/s	300GB/s	600 GB/s	200 GB/s
CUDA 코어	16,896	14,592	18,176	18,176	9,126	1280 × 4
Tensor 코어	989	456	568	568	288	40×4
시스템 인터페이스	PCIe Gen5 × 128GB/s	PCIe Gen5 × 128GB/s	PCIe Gen4 × 16 64GB/s			
최대 전력	600W	400W	350W	300W	150W	250W
Multi Instance GPU	최대 7 Instance	최대 7 Instance				

RTX(Quadro) GPU

제품	RTX 6000Ada	RTX 4000Ada	RTX A6000	RTX A5500	RTX A5000	RTX A4500	RTX A4000
메모리	48GB GDDR6 ECC 지원	20GB GDDR6 ECC 지원	48GB GDDR6 ECC 지원	24GB GDDR6 ECC 지원	24GB GDDR6 ECC 지원	20GB GDDR6 ECC 지원	16GB GDDR6 ECC 지원
대역폭	960 GB/s	280 GB/s	768 GB/s	768 GB/s	768 GB/s	640 GB/s	448 GB/s
CUDA 코어	18,176	6,144	10,752	10,240	8,192	7,168	6,144
Tensor 코어	568	192	336	320	256	224	192
RT 코어	142	48	84	80	64	56	48
시스템 인터페이스	PCIe Gen4 X 16 64GB/s	PCI Express 4.0 × 16	PCIe Gen4 X 16 64GB/s				
폼 팩터	Dual Slot 4.4"H X 10.5"L	69 × 178 (mm), Dual slot	Dual Slot 4.4"H X 10.5"L	Single Slot 4.4"H X 9.5"L			
최대 전력	300W	70w	300W	230W	230W	200W	140W

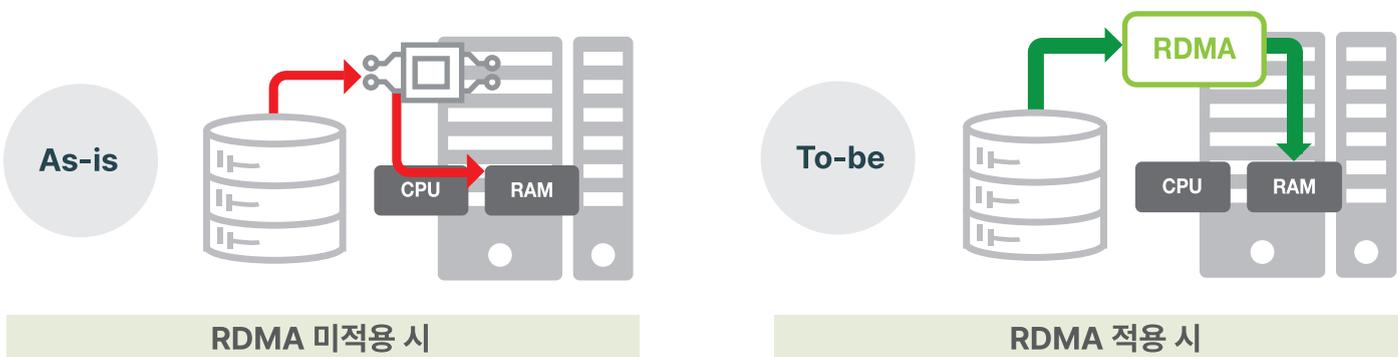
NVIDIA Network

01 엔드 투 엔드 인터커넥트 솔루션 제공 가능한 유일한 벤더



NIC, 스위치, 케이블, 트랜시버 자체 개발 및 턴키(Turn-key) 솔루션 제안 가능
데이터센터에 최적화된 솔루션 제공

02 RDMA, RoCE 성능 극대화 RDMA (Remote Direct Memory Access)



CPU를 통한 데이터 전송 (Onload)

CPU를 통하지 않고 Memory에 접근 가능(Offload)
CPU를 통하지 않기 때문에 CPU load가 낮고
데이터 속도가 빠름
더 빠른 데이터 속도 및 네트워크 내 컴퓨팅으로
더 높은 성능 및 용량 지원
RDMA는 인피니밴드 통신 기술

이더넷 스위치

네트워킹 라인업

SN3000 series 

SN3420	SN3510	SN3700	SN3700C
			
48-ports 25G + 12-ports 100G Ideal 1/10/25G aggregation for servers and storage 48×1/10/25GbE 12×100GbE Supports 24×50GbE, 12×40GbE, 96×1/10/25GbE 4.8Tbps	48-ports 50G + 6-ports 400G Ideal 1/10/25/50G aggregation for servers and storage 48×50GbE 6×400GbE Supports 6×400GbE 12×200GbE, 24×100GbE, 96×1/10/25/50GbE 9.6Tbps	32-ports 200G Ideal 100G/200G aggregation for servers and storage 32×200GbE Supports 64×100GbE, 32×40GbE, 128×1/10/25/50GbE 12.8Tb/s	32-ports 100G Ideal 100G aggregation for servers and storage 32×100GbE Supports 32×40GbE 128×1/10/25GbE 6.4Tb/s

* Supported using split cables

SN4000 series 

SN4600	SN4600C	SN4700	SN4800
			
64-ports 200G Ideal 200G aggregation for servers and storage 64×200GbE Supports 128×100GbE, 64×40GbE, 128×1/10/25/50GbE 25.6Tbps	64-ports 100G Ideal 100G aggregation for servers and storage 64×100GbE Supports 64×40GbE, 128×1/10/25/50GbE 12.8Tb/s	32-ports 400G Ideal 100G aggregation for servers and storage 32×400GbE Supports 64×200GbE, 128×100GbE, 64×40GbE, 128×1/10/25/50GbE 25.6Tb/s	128-ports 100G Ideal 400G aggregation for servers and storage 32×400GbE or 64×200GbE or 128×100GbE Supports 64×40GbE, 128×1/10/25GbE 25.6Tbps

* Supported using split cables

인피니밴드 스위치

네트워킹 라인업

Mellanox Quantum-1 HDR (200Gb/s) Switch

QM8700	QM8790	CS8500
		
<p>40-ports HDR</p> <ul style="list-style-type: none"> • 40 X HDR 100Gb/s ports in a 1U switch • 16Tb/s aggregate switch throughput • Up to 7.2 billion messages-per second • 130ns switch latency • Managed switch • 2 power supplies (1+1), hot swappable 	<p>40-ports HDR</p> <ul style="list-style-type: none"> • 40 X HDR 100Gb/s ports in a 1U switch • 16Tb/s aggregate switch throughput • Up to 7.2 billion messages-per second • 130ns switch latency • Unmanaged switch • 2 power supplies (1+1), hot swappable 	<p>800-ports HDR</p> <ul style="list-style-type: none"> • HDR X 800 ports in a 29U switch • 320Tb/s aggregate switch throughput • Ultra low switch latency

Mellanox Quantum-2 NDR (400Gb/s) Switch

QM9700	QM9790
	
<p>64-ports NDR</p> <ul style="list-style-type: none"> • 64 ports of 400Gb/s (NDR) over 32 OSFP cages • 51.2Tb/s aggregate switch throughput • 66.5 billion messages-per second • SHARP v3 • Managed switch • 2 power supplies (1+1), hot swappable 	<p>64-ports NDR</p> <ul style="list-style-type: none"> • 64 ports of 400Gb/s (NDR) over 32 OSFP cages • 51.2Tb/s aggregate switch throughput • 66.5 billion messages-per second • SHARP v3 • Unmanaged switch • 2 power supplies (1+1), hot swappable

어댑터 카드 (NIC)

네트워킹 라인업

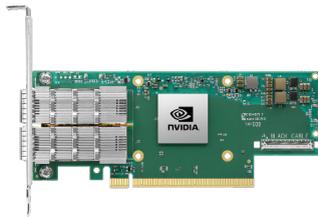


ConnectX-7

NVIDIA Quantum-2 InfiniBand 아키텍처 기반

ConnectX-7은 Low Latency, 400Gb/s 처리량, 혁신적인 NVIDIA 인-네트워크 컴퓨팅 가속화 엔진을 통해 슈퍼컴퓨터, 인공지능, 하이퍼스케일 클라우드 데이터센터에 필요한 확장성과 다기능 기술을 더욱 빠르게 제공합니다.

Form Factor		PCIe Standup Card		PCIe Standup - Socket Direct 16 + 16	
PCIe		Gen5.0 ×16	Gen5.0 ×16	2x Gen4.0 ×16 (x32)	
Port Speed		200Gb/s	400Gb/s	200Gb/s	400Gb/s
OPN (ports)	OSFP	MCX75310AAS-HEAT (1) MCX75210AAS-HEAT (1) option for 2×8 bifurcation	MCX75310AAS-NEAT (1) MCX75210AAS-NEAT (1) option for 2×8 bifurcation	MCX75510AAS-HEAT (1)	MCX75510AAS-NEAT (1)
	QSFP			MCX755105AS-HEAT (1) MCX755106AS-HEAT (2) VPI support	



ConnectX-6

NVIDIA Quantum InfiniBand 아키텍처 기반

HPC, 인공지능, 클라우드, 하이퍼스케일, 스토리지 플랫폼에서 효율성을 극대화하기 위한 고성능 NVIDIA 인-네트워크 컴퓨팅 가속화 엔진을 제공합니다.

Form Factor	PCIe Standup Card			PCIe Standup Card Socket Direct 8 + 8	PCIe Standup - Socket Direct 16 + 16	
PCIe	Gen4.0 ×16	Gen4.0 ×16	Gen4.0 ×16	Gen4.0 ×16	2x Gen3.0 ×16 (x32)	
Port Speed	100Gb/s	100Gb/s	200Gb/s	100Gb/s	100Gb/s	200Gb/s
OPN (ports)	MCX651105A-EDAT (1)	MCX653105A-ECAT (1) MCX653106A-ECAT (2)	MCX653105A-HDAT (1) MCX653106A-HDAT (2)	MCX653105A-EFAT (1) MCX653106A-EFAT (2)	MCX654106A-ECAT (2)	MCX654105A-HCAT (1) MCX654106A-HCAT (2)

트랜시버



QSFP56
HDR / 200GbE



QSFP112 & OSFP
NDR 400G



OSFP
Twin NDR 800G

케이블



Direct Attach Cables (DACs)
• 400Gb/s, NDR(400G)
• 200Gb/s, HDR(200G)



Active Optical Cables (AOCs)
• 400Gb/s
• 200Gb/s, HDR(200G)



Passive Optical Cables (POCs)
• NDR(400G)



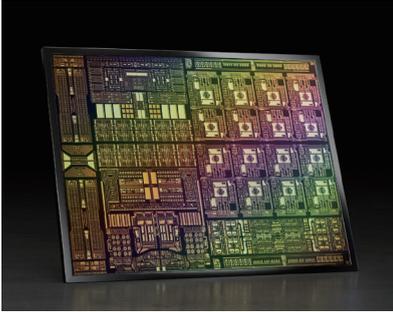
DAC 분기 케이블



AOC 분기 케이블

NVIDIA DPU

01 NVIDIA BLUEFIELD 3



3세대 DPU(data processing unit)인 NVIDIA® BlueField® -3는 소프트웨어 정의 및 하드웨어 가속 기능을 사용하여 클라우드, 코어 데이터 센터, 엣지 인프라 등 모든 위치에서 효과적이며 효율적인 IT 인프라를 구축할 수 있도록 지원합니다.

400Gb/s 이더넷 또는 NDR 400Gb/s 인피니밴드 네트워크 연결을 제공하는Blue-Field-3 DPU는 데이터 센터 성능, 시스템 운영 효율 및 보안을 대대적으로 개선하는 방식으로 소프트웨어 정의 네트워킹, 스토리지, 보안 및 관리 기능을 오프로드 기반으로 가속 및 분리시킬 수 있습니다.

또한 I/O 경로에서 다양한 프로그램 엔진을 가속화 시키는 기능과 강력한 컴퓨팅 성능을 제공하는 BlueField-3은 가장 많은 리소스를 요하는 어플리케이션의 인프라 요건을 충족시키고, NVIDIA DOCA™ 소프트웨어 프레임워크를 통해 이전 소프트웨어 어와도 완벽한 호환성을 제공할 수 있으며 기존 컴퓨팅 환경에 보안 기능을 강화하고 가속화된 성능으로 가상의 프라이빗 클라우드로 전환시켜, 안전한 멀티-테넌트 환경에서 어플리케이션 워크로드를 처리할 수 있게 해줍니다.

02 핵심 소프트웨어- 정의, 하드웨어-가속 어플리케이션



클라우드 네트워킹

클라우드 오버레이, SDN 가속
NAT, 로드 밸런서, NFV,
비디오 스트리밍



스토리지

NVMe-oFTM(NVMe™ over
Fabrics), NVMe/TCPTM,
엘라스틱 스토리지(elastic
storage), 하이퍼 컨버지드
인프라(HCI), 암호화, 데이터
무결성, 데이터 중복제거(data
deduplication),
압축해제(decompression),
EC(erasure coding)/RAID



보안

분산 차세대 방화벽,
IDS/ IPS, RoT
(Root of Trust),
마이크로 분할
(microsegmentation),
DDOS 방지



HPC /AI

클라우드-네이티브
슈퍼컴퓨팅,
멀티-테넌시(multi-tenancy)
및 보안, 통신가속



통신 & 엣지

클라우드 RAN,
가상화 엣지 게이트웨이,
VNF 가속,
엣지 마이크로서버

03 라인업

Model	BRACKET SIZE	CPU	M/M	PCI	SPEED	Port	Embedded OPTION
900-9D3B4-00CC-EA0	Tall Bracket	E-series 8 Arm cores	16GB	PCIe Gen4.0 ×16	100GbE/HDR100	Dual	Crypto Enabled
900-9D3B4-00EN-EA0					400GbE/NDR	Single	Crypto Enabled
900-9D3D4-00EN-HA0					400GbE/NDR	Single	Crypto Enabled
900-9D3B4-00PN-EA0					400GbE/NDR	Single	Crypto Disabled
900-9D3B4-00SC-EA0					100GbE/HDR100	Dual	Crypto Disabled
900-9D3D4-00NN-HA0					400GbE/NDR IB	Single	Crypto Disabled
900-9D3B6-00CC-AA0		P-Series 16 Arm cores	32GB	PCIe Gen5.0 ×16 (x16 PCIe extension option)	100GbE/HDR100	Dual	Crypto Enabled
900-9D3B6-00CV-AA0					200GbE/NDR200	Dual	Crypto Enabled
900-9D3B6-00SC-AA0					100GbE/HDR100	Dual	Crypto Disabled
900-9D3B6-00SV-AA0					200GbE/NDR200	Dual	Crypto Disabled
900-9D3C6-00CV-DA0					200GbE/NDR200	Dual	Crypto Enabled Secure Boot

Business Partner & Customer Base

다양한 국내외 파트너사 & 고객사 레퍼런스 보유

국내외 파트너사



주요 고객 현황

Telco 및 방송사



기업



금융/CAMPUS/공공





아이크래프트 Do it AI 는 (주)아이크래프트 내 AI 사업 본부로서
기업이 AI 프로젝트를 수행할 수 있도록 AI 인프라의 초기 설계부터
안정적인 최종 구축까지 맞춤형 One-stop 서비스를 제공합니다



KOSDAQ
코스닥상장법인



아이크래프트(주)

서울시 강남구 테헤란로 44길 8, 아이콘 역삼빌딩 7F, 8F (우)06210 | T. 02-541-0474
www.icraft21.com | www.do-it-ai.com | E. ai_sales@icraft21.com
COPYRIGHT© 2024 iCRAFT Co. Ltd. All rights reserved.



아이크래프트 DO-IT-AI 홈페이지에서
더 많은 정보를 알아보세요!
<http://www.do-it-ai.com/>