생성형 AI를 위한 Full Stack Validated Design

2024.05

전략기술 사업부 유채호

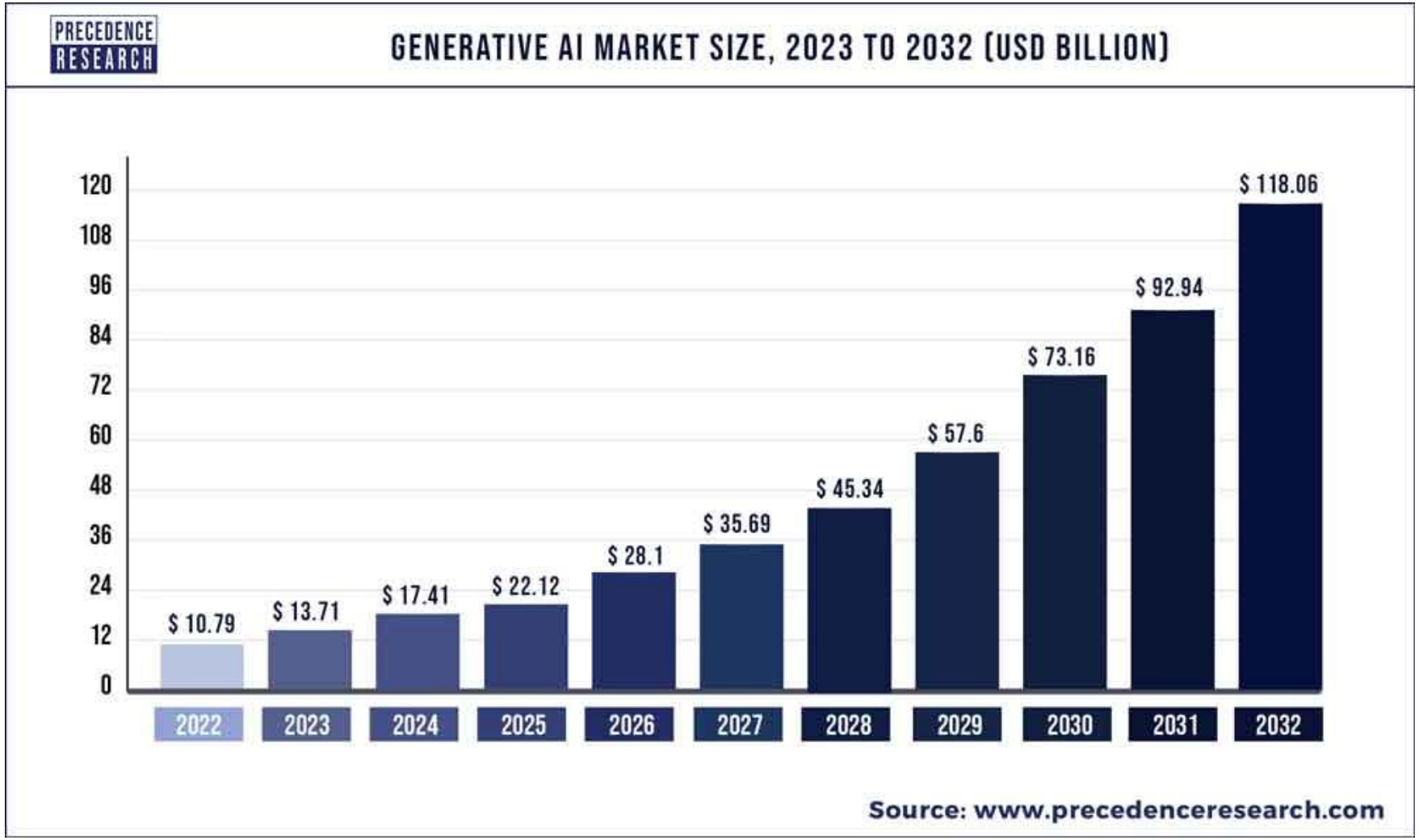
MAYMUST

목차

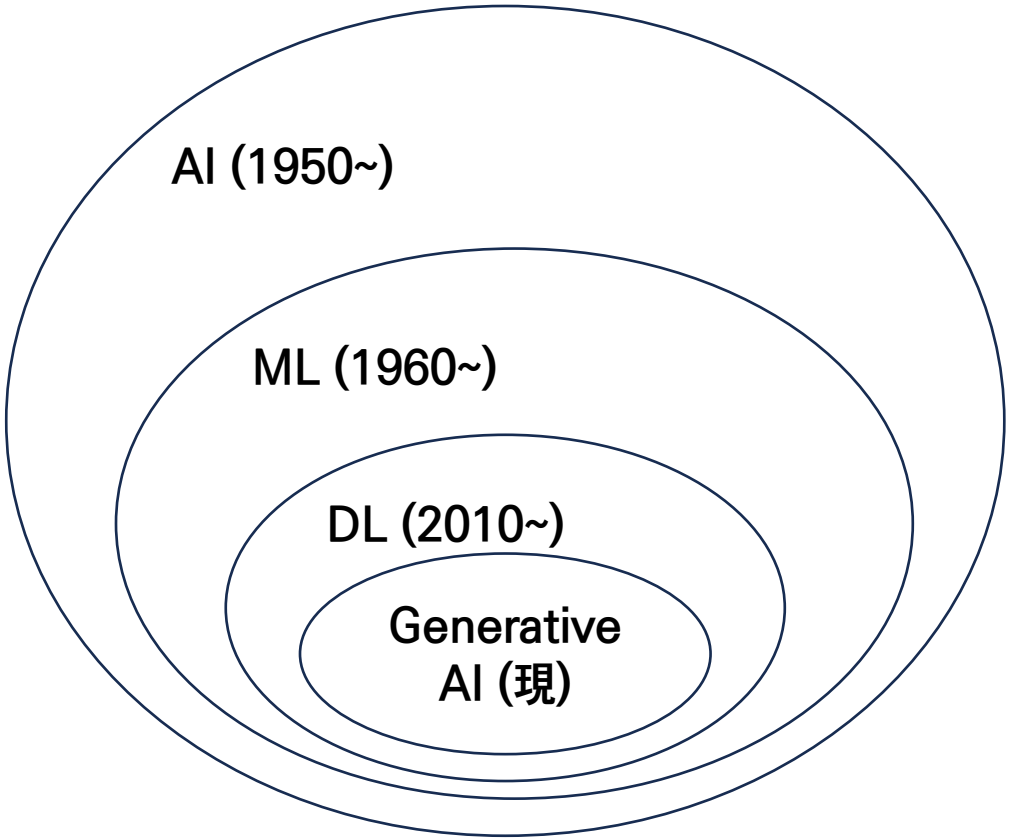
Introduction to Gen AI

1. Introduction (1/2)

생성형 AI (Generative AI) 시장 추이



1. Introduction (2/2)



● 인공지능(AI) History

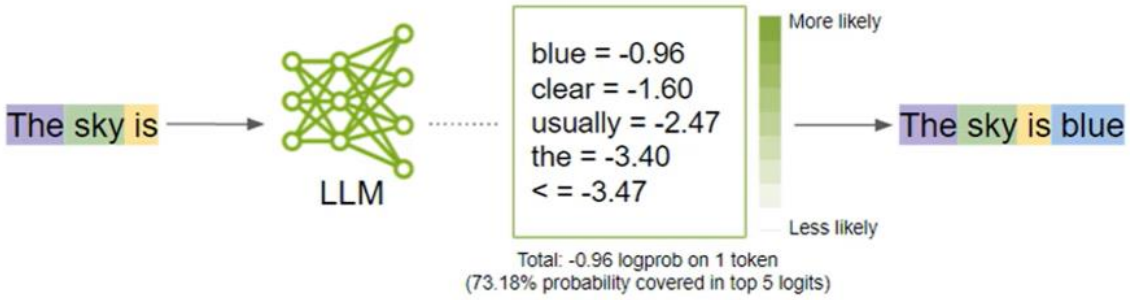
1. AI – 규칙기반 시스템 (1950년대~1960년대)
2. ML – 머신 러닝 (1960년대 ~1990년대)
3. DL – 딥 러닝 (2010년대 ~ 현재)
4. Gen AI – 생성형 AI (현재)

- 2017 : “transformer” 모델에 의한 LLM 혁명 시작
  - token 처리를 통한 DL architecture 특화
  - self-attention 활용. “Attention is All you Need” 논문
- 2018 : GPT 개발 (Open AI)
- 2023 : Chat GPT3

2. Generative AI 배경 및 개념

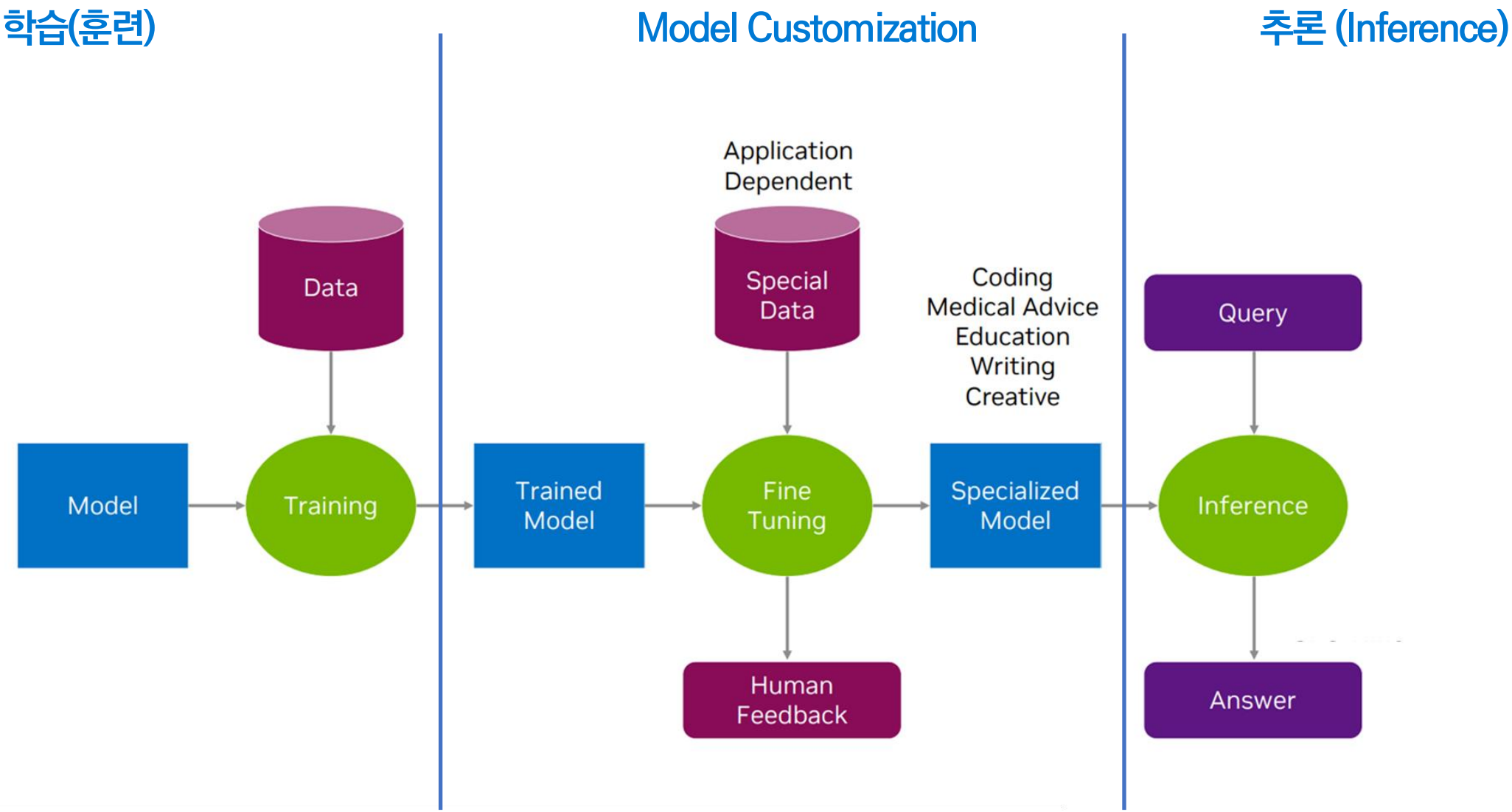
● 정의 및 개요

- **Generative AI** – 인간이 명시적으로 프로그래밍하지 않고 기존 예제와 스타일&구조가 유사한 콘텐츠(ex. 이미지, 텍스트, 또는 오디오)를 생성할 수 있는 모델을 구축하는 인공지능의 한 분야
- **Transformer** – 2017년 Vaswani 논문에서 소개. 단어 문맥 관계를 학습
- **GPT (Generative Pretrained Transformer)** – Open AI 가 개발한 Transformer 기반의 LLM
- **LLM (Large Language Model)** – 일반적으로 자연어 텍스트를 이해, 처리 및 생성할 수 있는 딥 러닝 기술을 사용하여 광범위한 데이터 세트에서 훈련된 고급 유형의 AI 모델
- **Foundation Model** – 레이블이 지정되지 않은 광범위한 데이터 집합에 대해 훈련된 대규모 인공 지능 모델. (GPT, llama2, etc.)



- Transformer models built with unsupervised learning proved to be effective next-token predictors
- Foundation models: Trained on massive unlabeled datasets and can be tuned to specialized applications with comparatively few examples
- Large Language Model: Scaled-up architectures that can accomplish language-related tasks like summarizing, translating, or composing new content

3. Type of Workload (워크로드 유형)



4. Benefits and Use cases

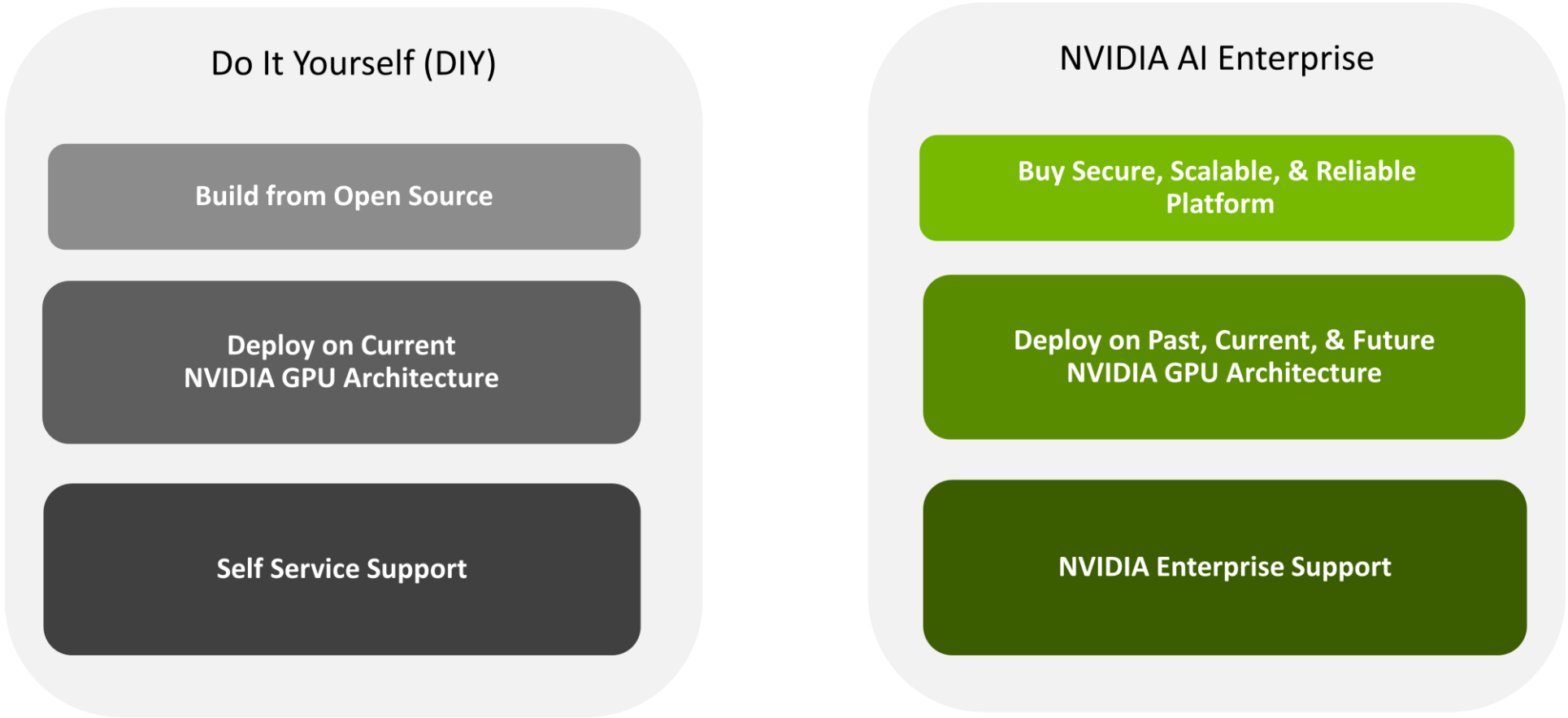
● Generative AI 의 장점

- 생산성 향상 (Improved productivity)
- 향상된 고객 경험 (Enhanced customer experience)
- 더 나은 의사 결정 (Better decision-making)
- 비용 절감 (Cost savings)
- 혁신 증대 (Increased innovation)
- 경쟁 우위 (Competitive advantage)

● Generative AI 의 Use Case

- 고객 서비스 (Customer Service)
- 콘텐츠 만들기 (Contents creation)
- 제품 디자인 (Product Design)
- 교육 (Education)
- 사기 탐지 (Fraud detection)
- 의료 (Healthcare)
- 게임 (Gaming)
- 소프트웨어 개발 (Software development)

5. Two Ways to Build AI Platform

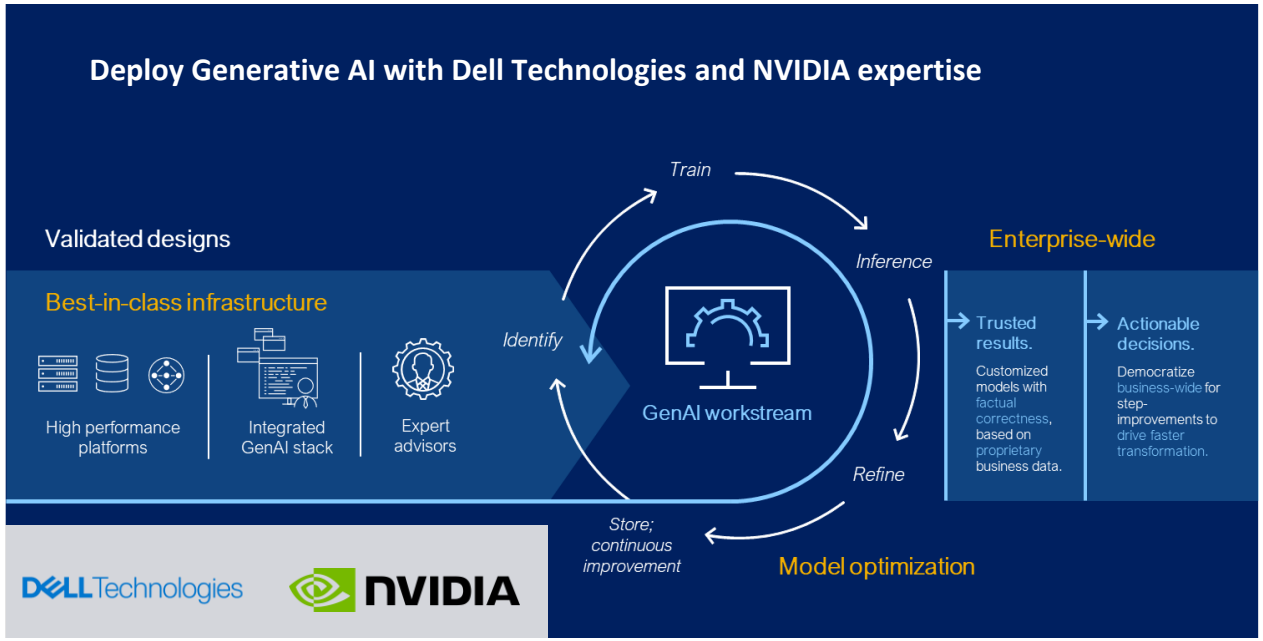


DELL x NVIDIA Project Helix

Project Helix

Blueprint for on–premises generative AI

Project Helix is a full–stack solution that enables enterprises to create and run custom AI models built with the knowledge of their business

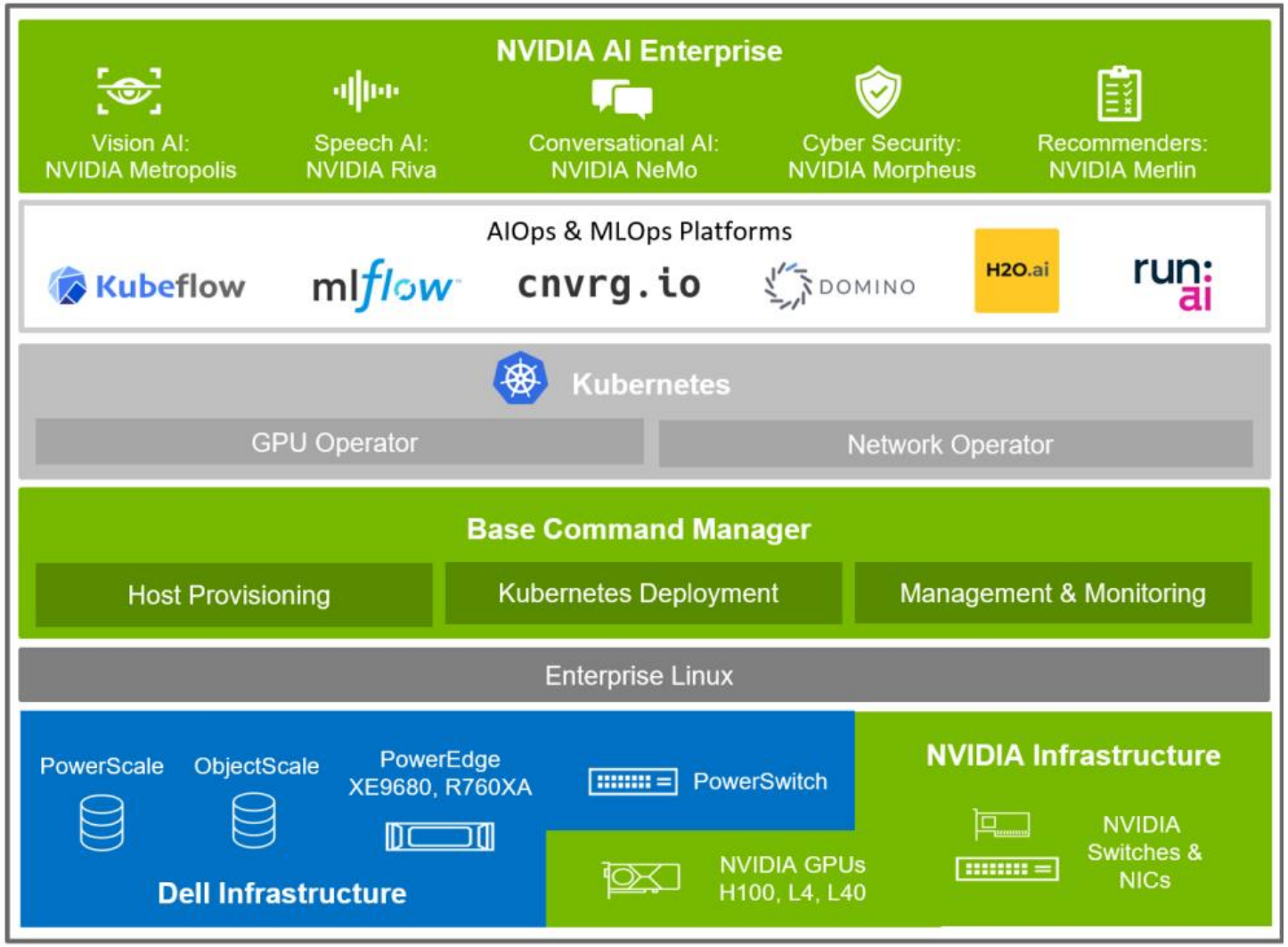


Dell + NVIDIA 의 장점

- Dell (HW) + NVIDIA (SW) 풀스택 Gen AI 솔루션 제공
- 가치 창출 시간을 단축하는 검증된 설계(아키텍처) 제공
- Sizing 및 scaling 정보 제공
- 지속적으로 개선 가능한 맞춤형 Gen AI model 제공

Dell and NVIDIA Solution Architecture

Scalable, Modular, Secure, and High-performance Solution



Dell Validated Designs based on the Project Helix initiative includes:

•**Generative AI training and AI inferencing Computing:** Dell PowerEdge servers, for example, the PowerEdge XE9680 and PowerEdge R760xa, which are optimized to deliver generative AI performance for Project Helix solutions.

•**Infrastructure:** When NVIDIA® H100 Tensor Core GPUs and NVIDIA Networking are combined with Dell servers, this will form the Project Helix infrastructure backbone.

•**Data Storage:** Customers can add Dell PowerScale and Dell ECS Enterprise Object Storage for scalable unstructured data storage

•SW Stack & Tools: NVIDIA AI Enterprise, including NVIDIA NeMo, and Base Command Essential
+ Dell OpenManage Enterprise, Dell OpenManage Enterprise Power Manager, Dell CloudIQ

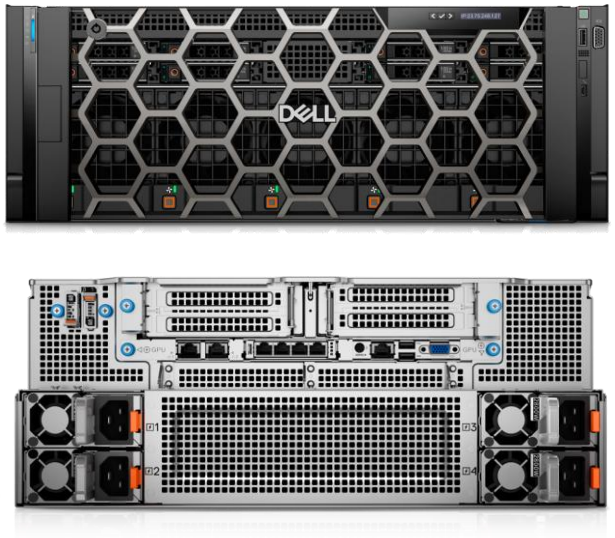
And, 3rd party MLOps Platform

Dell PowerEdge Server (for NVIDIA GPU)

XE9680

8 GPU

XE8640



4 GPU

R760xa

4 GPU

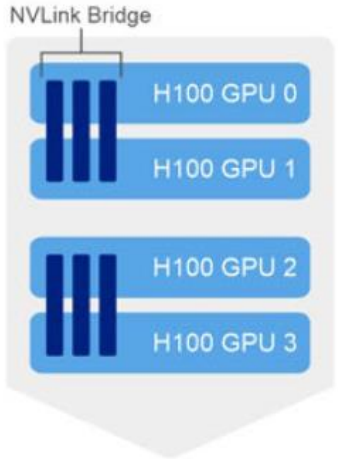
NVIDIA GPUs – Technical specification and use case

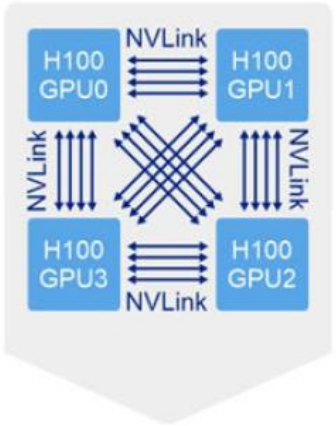
	NVIDIA H100 SXM GPU	NVIDIA H100 PCIe GPU	NVIDIA L40 PCIe GPU
Supported latest PowerEdge servers (and max # of GPUs)	PowerEdge XE9680 (8) PowerEdge XE8640 (4)	PowerEdge R760xa (4) PowerEdge R760 (2)	PowerEdge R760xa (4) PowerEdge R760 (2)
GPU Memory	80 GB	80 GB	48 GB
Form factor	SXM	PCIe	PCIe
Multi-instance GPU support	Up to 7 MIGs	Up to 7 MIGs	None
Max thermal design power (TDP)	700 W	350 W	300 W
NVIDIA AI Enterprise	Add-on	Included with H100 PCIe	Add-on
Use cases	Generative AI training Large scale distributed training	Discriminative/Predictive AI Training and Inference Generative AI Inference	Small scale AI Visual computing Discriminative/ Predictive AI Inference

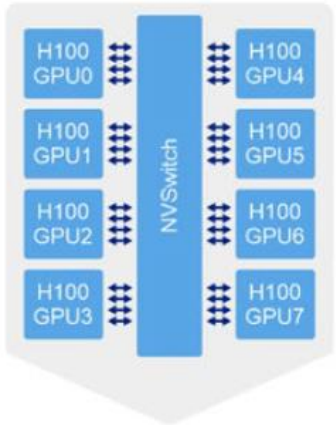
https://resources.nvidia.com/en-us-tensor-core/nvidia-tensor-core-gpu-datasheet

NVIDIA GPU Connectivity in PowerEdge servers

GEN AI 목적에 맞춤화된 여러 GPU 옵션을 갖춘 Dell PowerEdge 서버
NVLink 와 NVSwitch는 대규모 AI 응용 프로그램에 이상적인 고대역폭 (900 GB/s) 및 낮은 지연 시간의 데이터 전송 용이







DELL PowerEdge R760xa
NVIDIA L40S PCIe GPU

NVLink 미지원

DELL PowerEdge R760xa
NVIDIA H100 PCIe GPU

NVLink Bridge

DELL PowerEdge XE8640
NVIDIA H100 SXM5 GPU

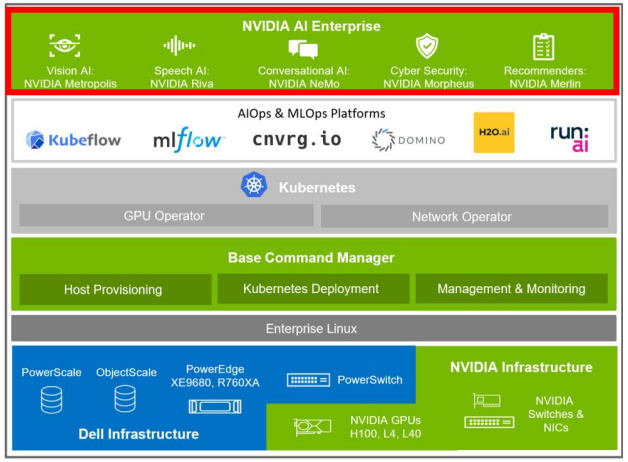
4세대 NVLink

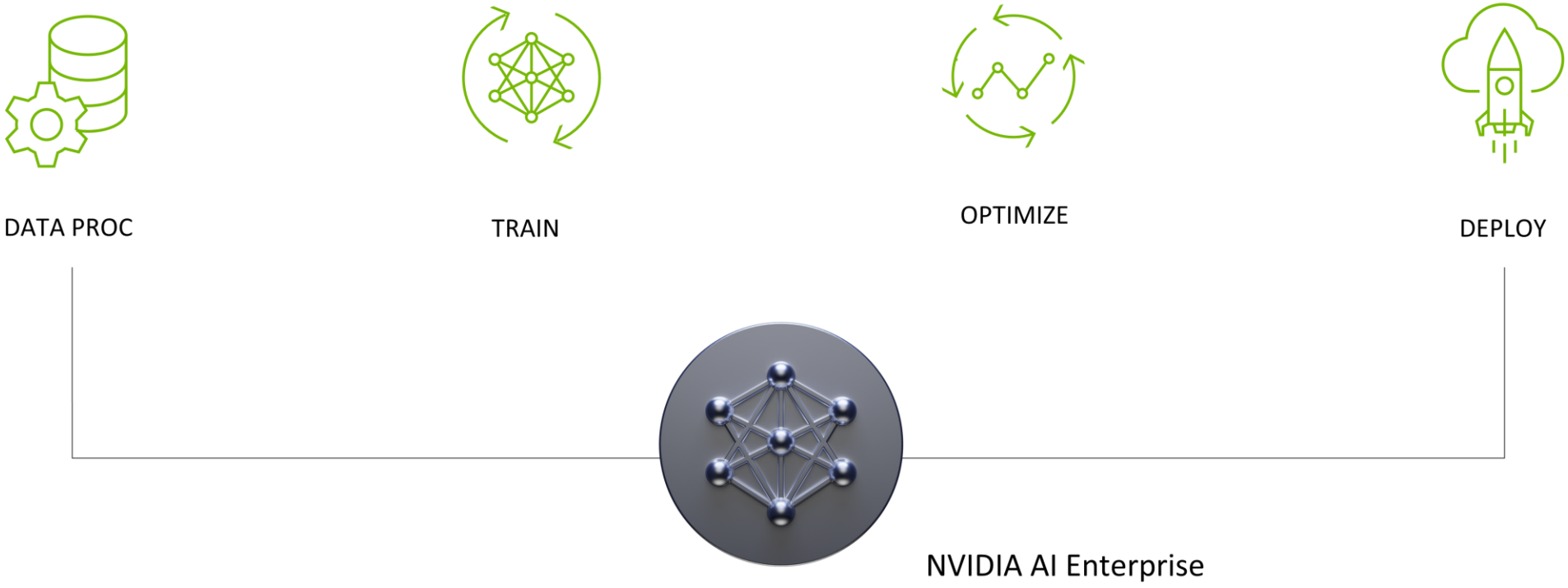
DELL PowerEdge XE9680
NVIDIA H100 SXM5 GPU

4세대 NVLink + 3세대 NVSwitch

MAYMUST

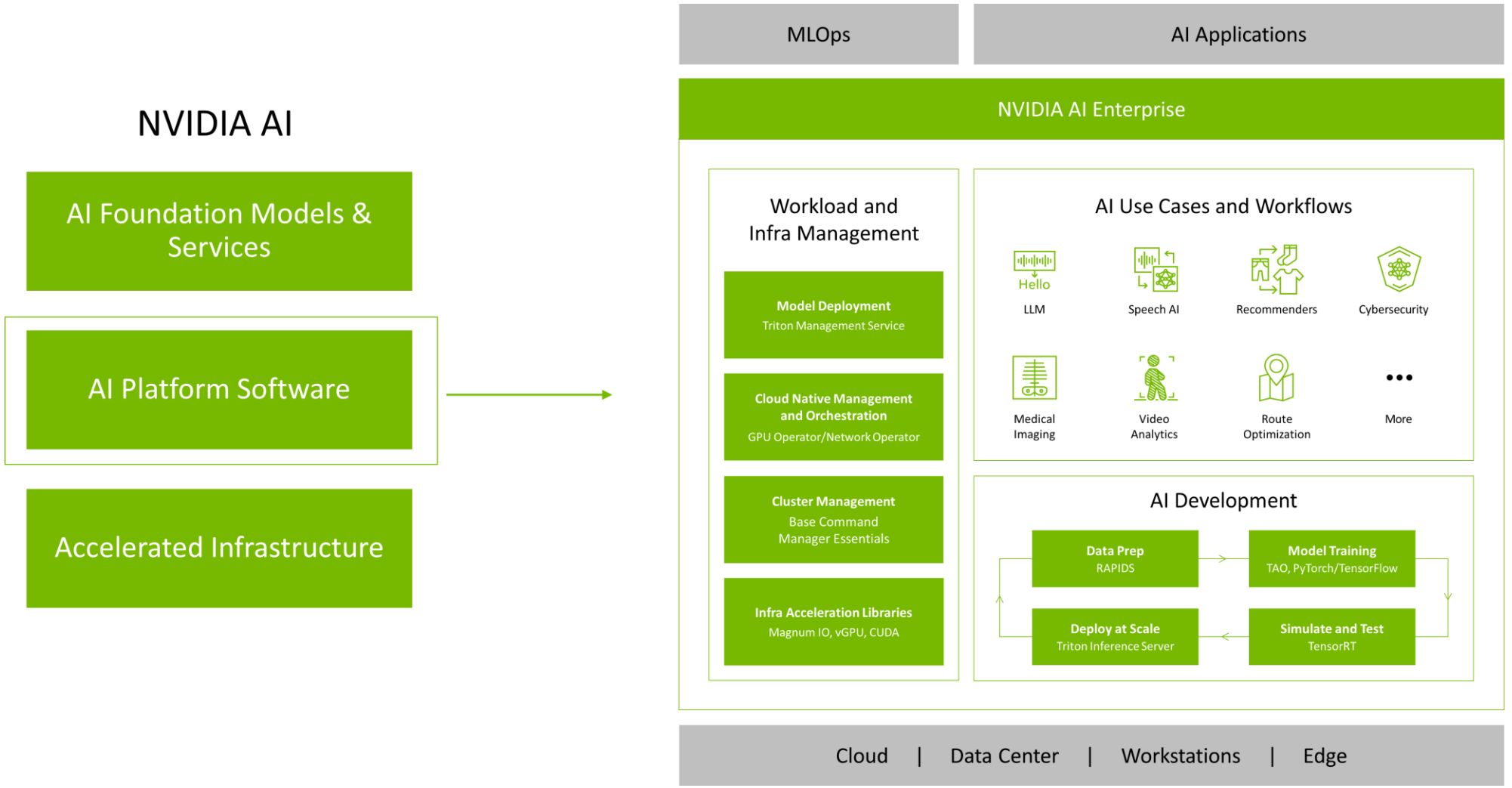
Software Development Platform for the AI Pipeline





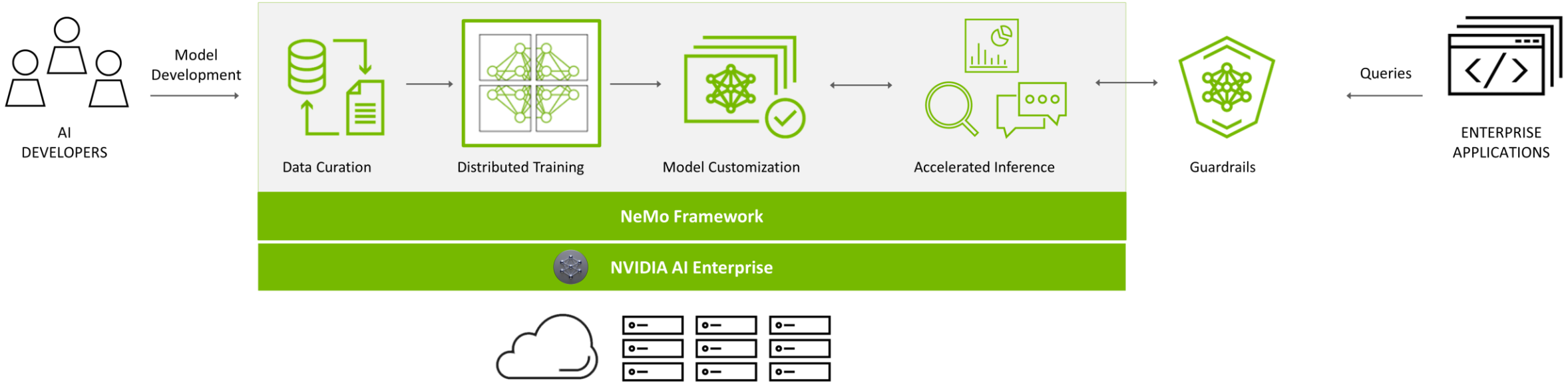
NVIDIA AI Enterprise

End to end AI Software



NVIDIA NeMo for Custom LLMs

- End-to-end, cloud-native framework to build, customize and deploy generative AI models



Multi-Modality

Build language, image, generative AI models

Data Curation at Scale

Extract, deduplicate, filter info from large unstructured data @ scale

Optimized Training

Accelerate training and throughput by parallelizing the model and the training data across 1,000s of nodes.

Model Customization

Easily customize with P-tuning, SFT, Adapters, RLHF, AliBi

Deploy at Scale

Run optimized inference at-scale anywhere

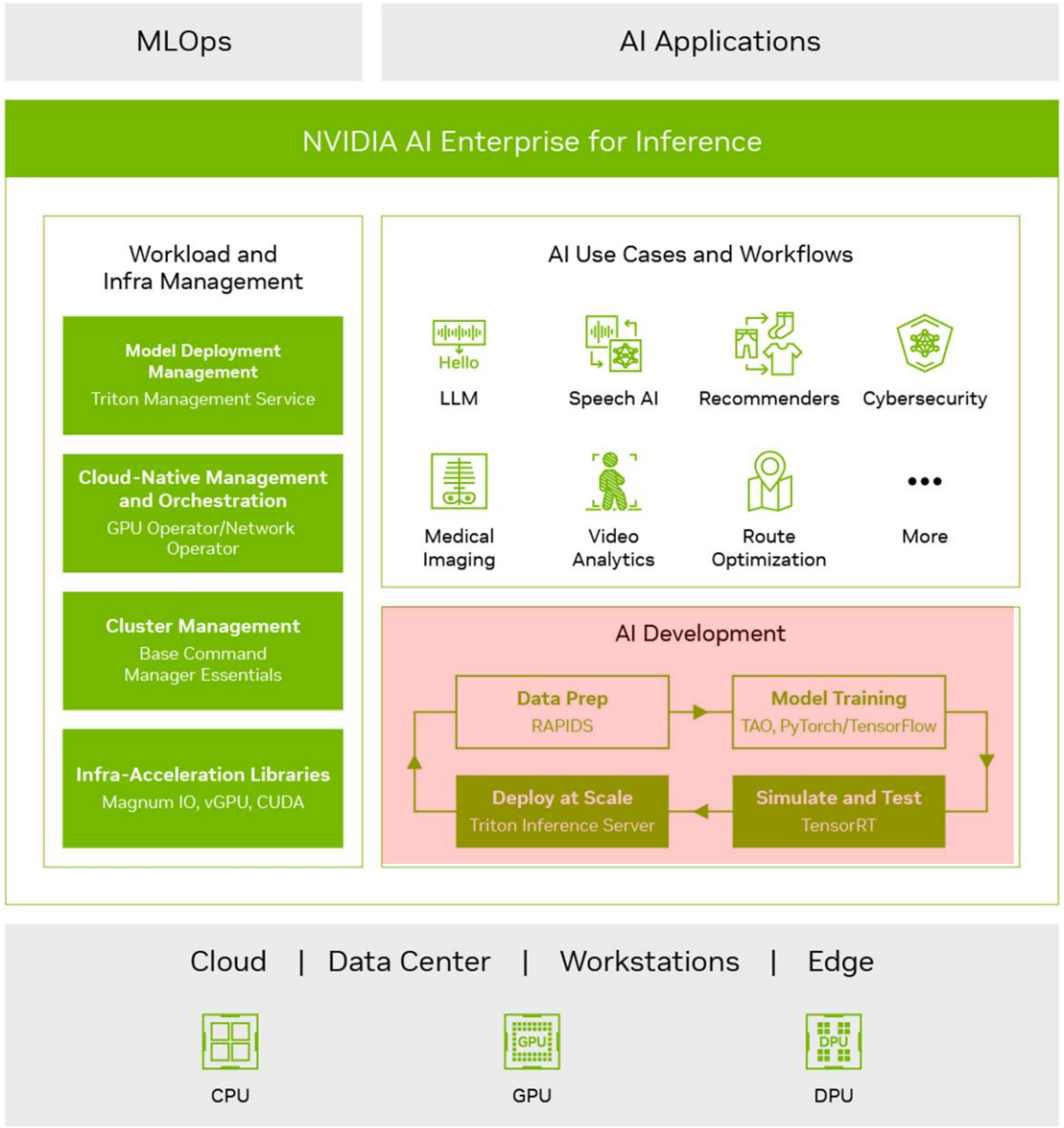
Guardrails

Keep applications aligned with safety and security requirements using NeMo Guardrails

Support

NVIDIA AI Enterprise and experts by your side to keep projects on track

NVIDIA AI Inference Platform for Enterprise



Triton 추론 서버

– standardizes AI model deployment and execution and delivers fast and scalable AI in production

- Any framework, any model, any platform, any query type, any deployment location, high performance

TensorRT

– SDK for high-performance DL inferencing, delivering low latency and high throughput for inference applications.

- **TensorRT-LLM** for the latest Large Language Models

Triton Management Service*

– enterprise control plane for Triton

- Simplified deployment, Resource Maximization, Monitoring & Self-healing

*Exclusive with an NVIDIA AI Enterprise subscription

MAYMUST

NVIDIA AI Workflows

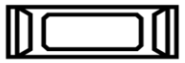
Prepackaged reference application to rapidly automate your business with AI

Generative AI Knowledge Base Chatbot	Intelligent Virtual Assistant	Audio Transcription	Digital Fingerprinting Threat Detection	Spear Phishing Detection	Route Optimization	Next Item Prediction
 **						
Generate accurate real-time responses from the company's knowledge base	Engaging contact center assistance 24/7 for lower operational costs	World-class, accurate transcripts based on GPU-optimized models	Cybersecurity threat detection and alert prioritization to identify and act faster	Use generative AI to improve the detection of spear phishing emails	Vehicle and robot routing optimization to reduce travel times and fuel costs	Personalized product recommendations for increased customer engagement and retention

NVIDIA AI Enterprise

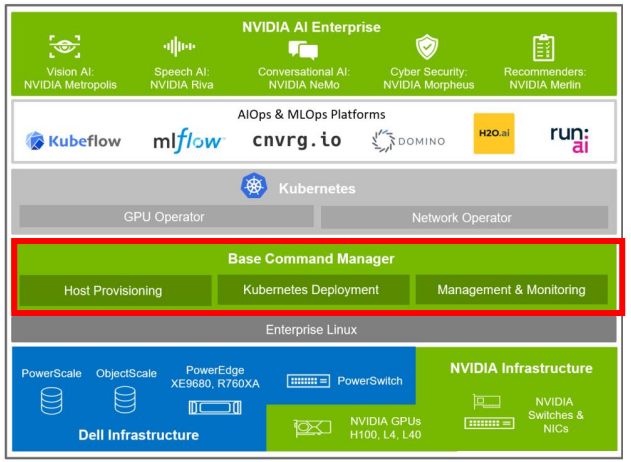
 Cloud

 Data Center

 Edge

 Embedded

Introducing Base Command Manager Essentials



Purpose-built for Enterprise AI Infrastructure Management

Infrastructure Provisioning

안전하고 신뢰할 수 있는
최신 AI 인프라 유지

Workload Management

Data scientist에 필요한
tool 및 리소스를 손쉽게 제공

Resource Monitoring

정보에 입각한 의사결정을
위한 자세한 통찰력 확보

Cloud

Data Center

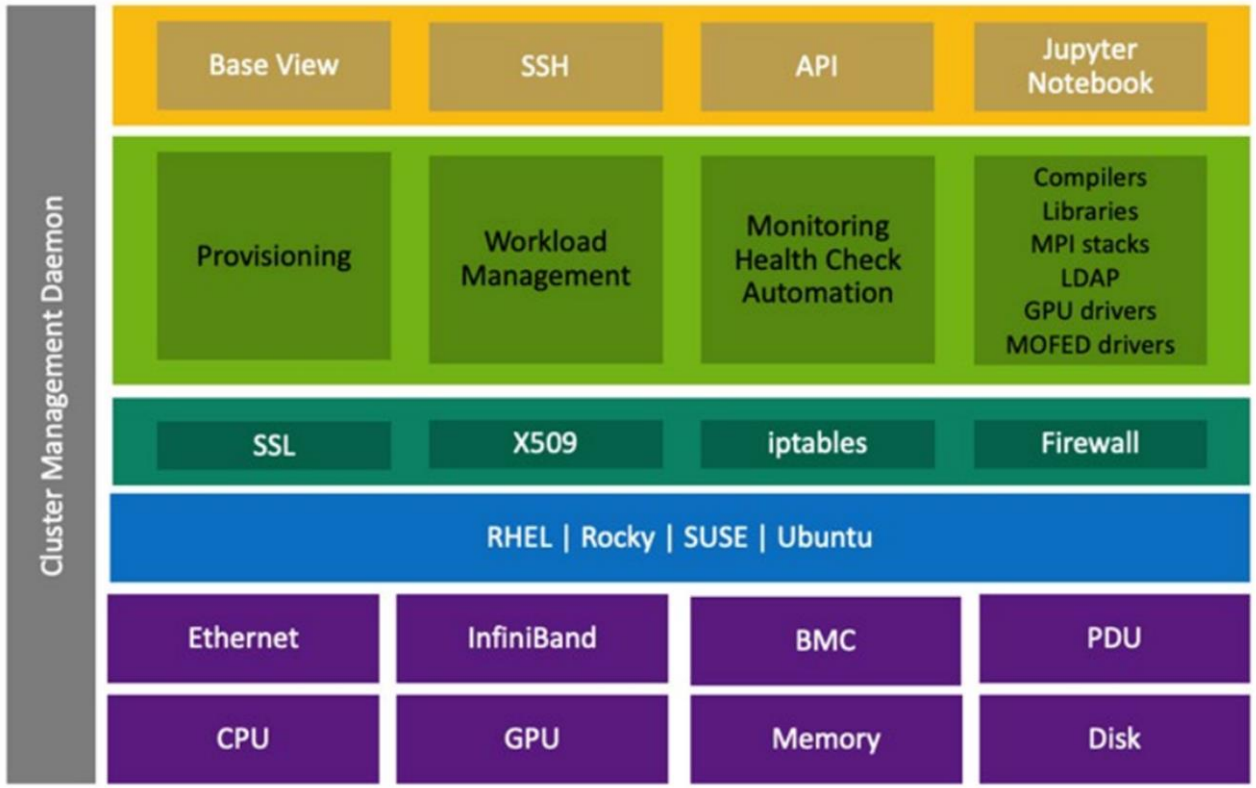
Edge

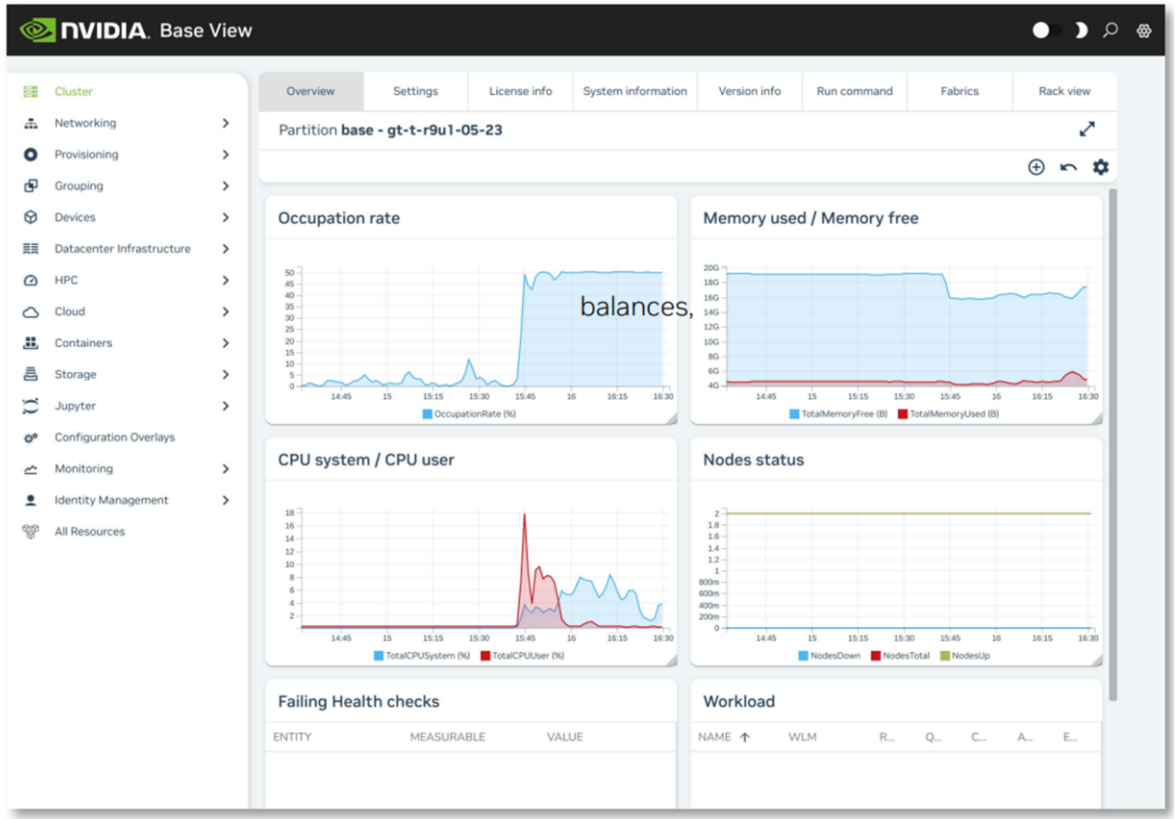
*Exclusively available with NVIDIA AI Enterprise

Base Command Manager Essentials*

Purpose-built for Enterprise AI Infrastructure Management

*Exclusively available with NVIDIA AI Enterprise

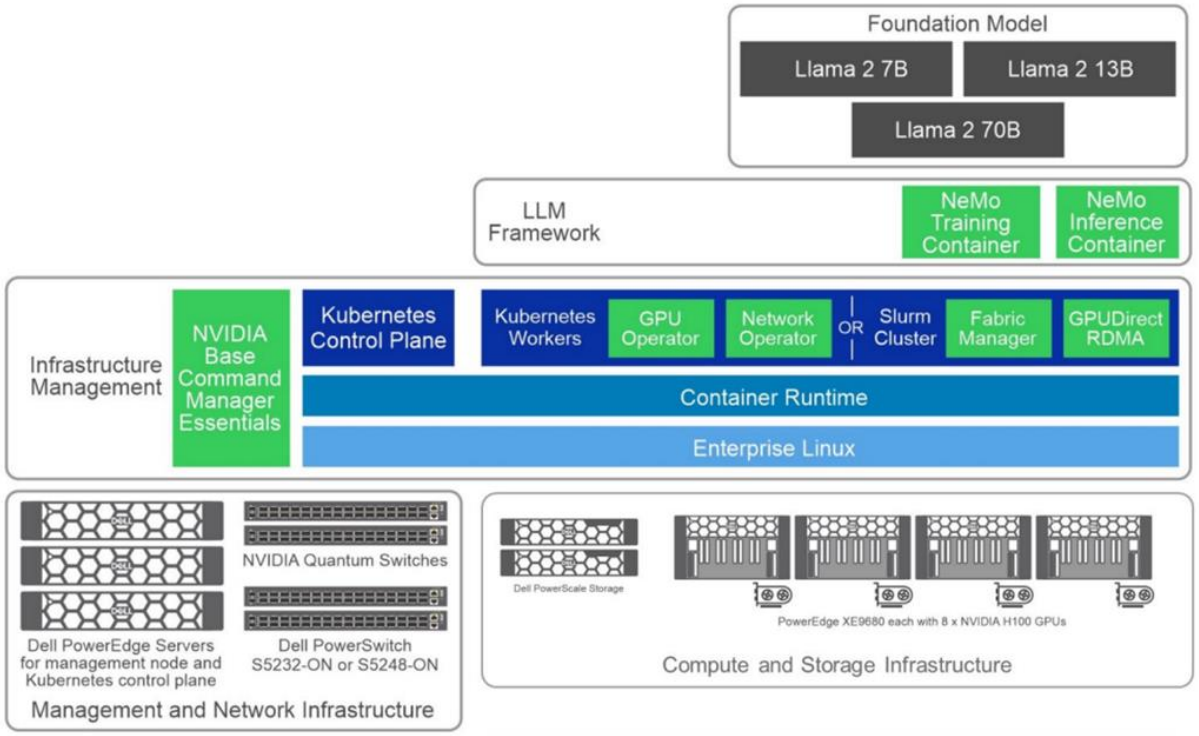


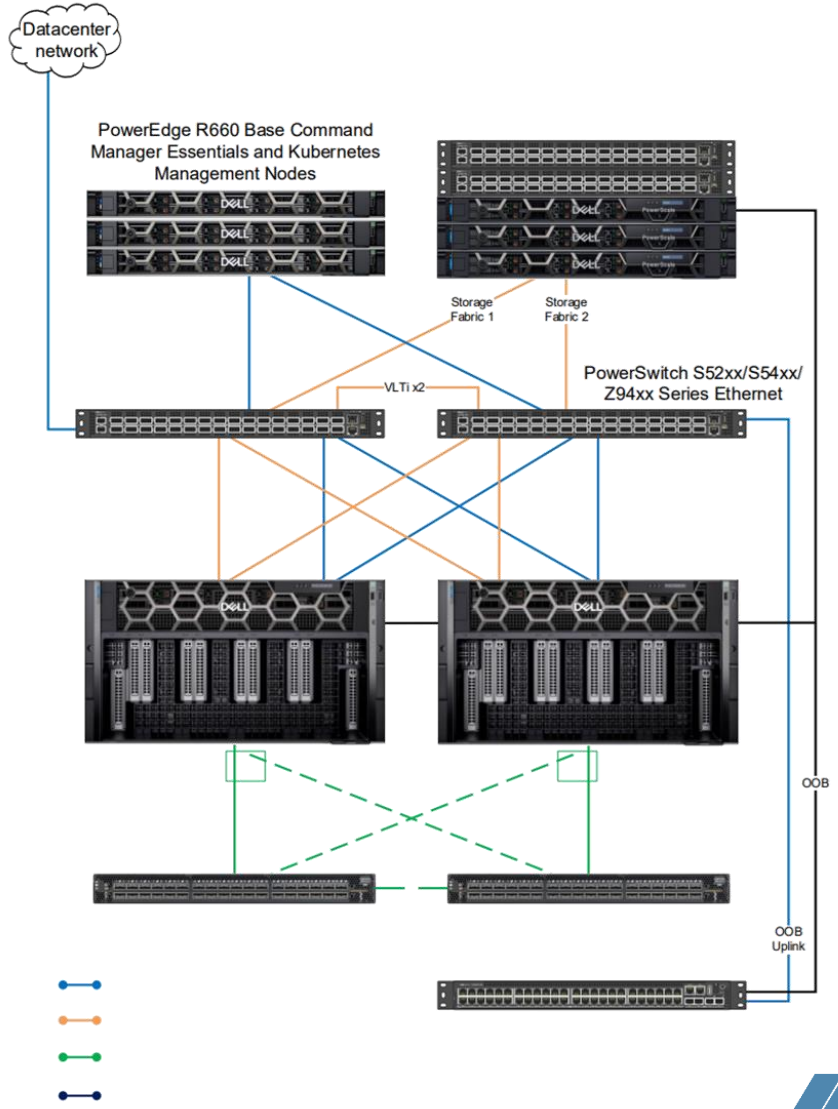


Base Command Manager comes with every DGX system.
For other systems, Base Command Manager Essentials is included with NVIDIA AI Enterprise

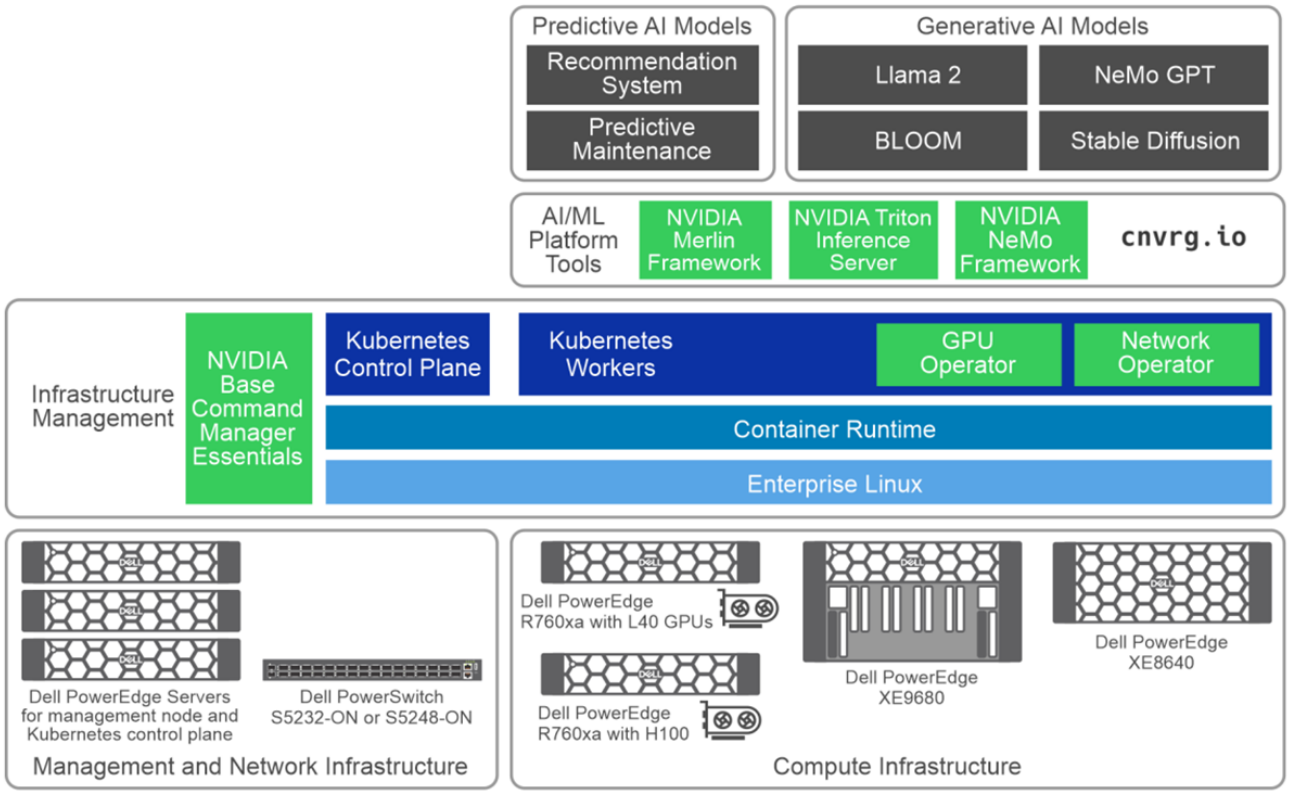
MAYMUST

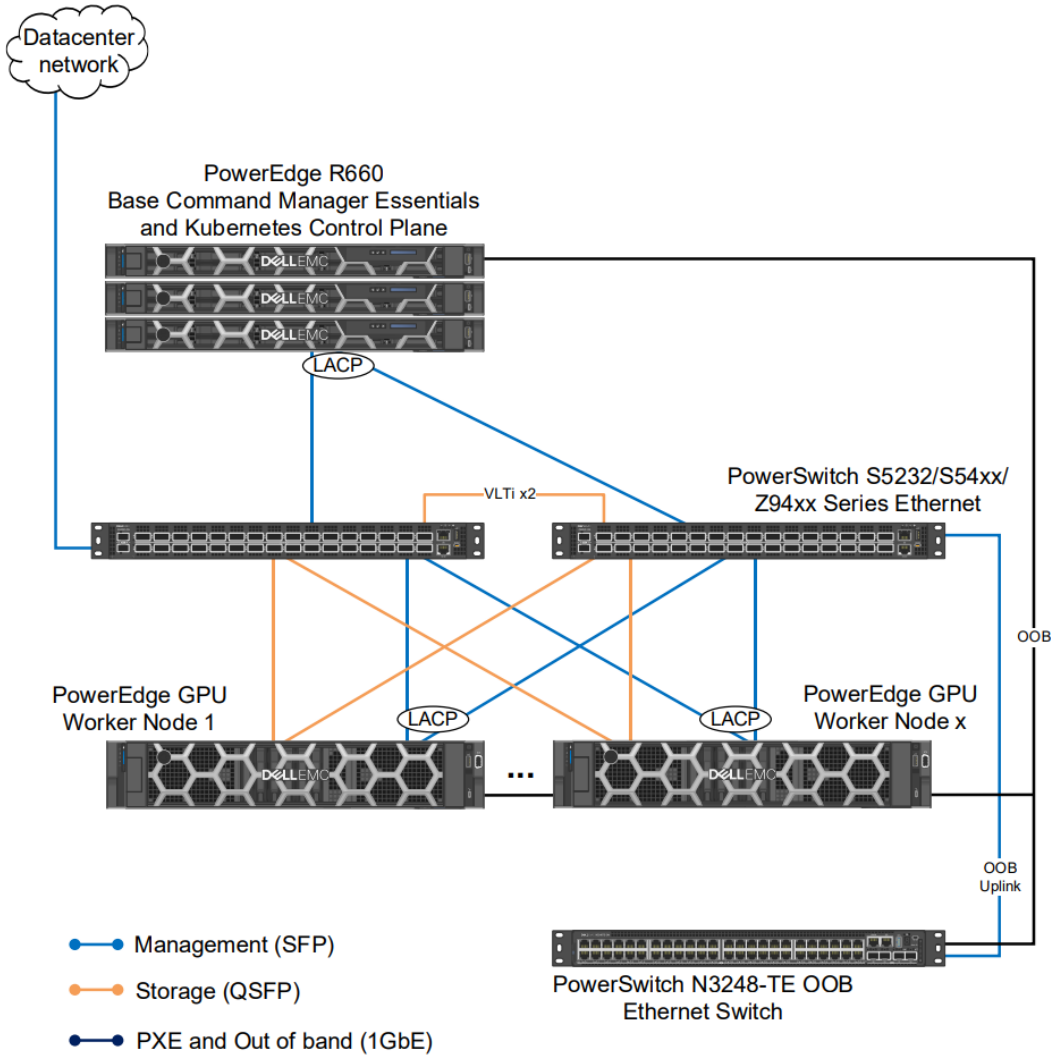
Reference architecture – LLM Customization





Reference architecture – Inferencing



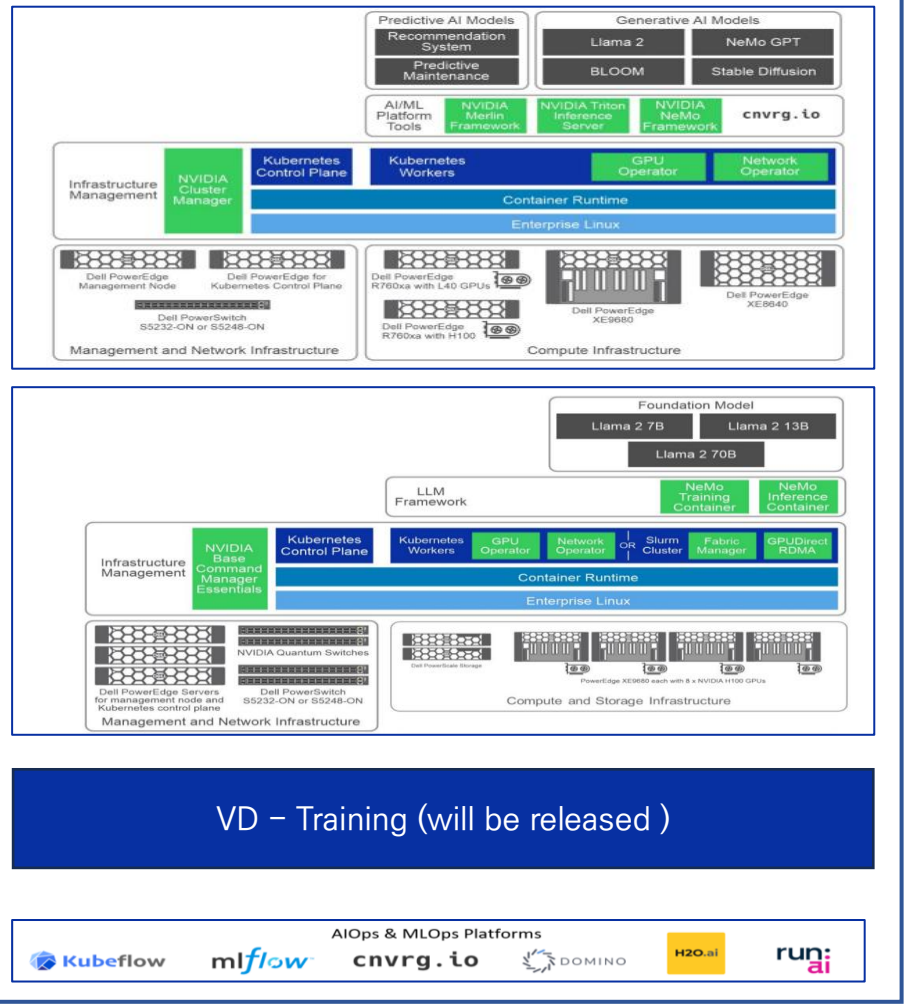


Project Helix 데모 센터

Project Helix 데모센터 운영 방안

Project Helix 데모 센터를 구축하여 AI 맞춤형 데모 및 교육에 활용

(Generative AI in the Enterprise – Model Customization, Reference architecture)



Dell Validated Designs
구축 및 검증

[GenAI Validated Designs]
Model Training, Customization, and Inferencing

SW Stack & Tools 구축
검증 및 테스트

NVIDIA AI Enterprise, NVIDIA NeMo, Base Command Essential 등 주기적으로 업데이트 되는 SW Stacks & Tools

고객 Use-Case 프로토타입/
PoC 지원 및 교육

고객 Use-Case 프로토타입/ PoC 진행

(3rd party) MLOps 플랫폼
정합 및 검증

국내(외) MLOps 등 파트너사들의 솔루션 정합 검증을 통한 에코시스템 확대

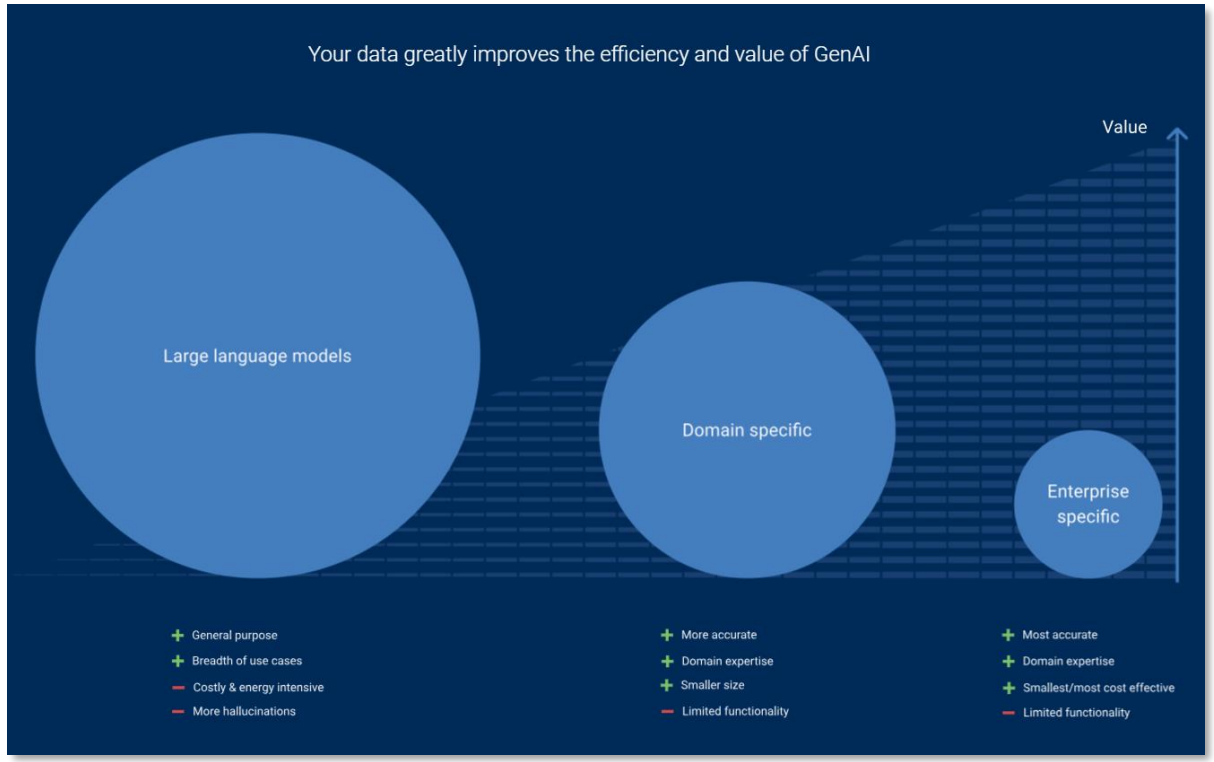
AI Infra (HW & SW) 구축 및 검증 테스트

Project Helix 기반의 GPU Cluster를 구축하고 성능을 검증

작업 시나리오	작업 참고 URL
Project Helix 기반 GPU Cluster Systems 구축 using Base Command Manager Essential	• https://infohub.delltechnologies.com/en-US/t/design-guide-generative-ai-in-the-enterprise-model-customization/ • https://infohub.delltechnologies.com/en-US/t/design-guide-generative-ai-in-the-enterprise-inferencing/ • https://docs.nvidia.com/base-command-manager/index.html#product-manuals
SW 설치 검증 및 GPU Burn/DCGM 검증	• https://github.com/wilicc/gpu-burn • https://github.com/NVIDIA/DCGM - GPU Driver, Fabric-Manager, CUDA Toolkit, …
NCCL (Single/Multi) 테스트	• https://github.com/NVIDIA/nccl
MLPerf (Single/Multi) 테스트	• https://github.com/mlcommons/training_results_v3.1/tree/main/NVIDIA/benchmarks/gpt3/implementations/pytorch
HPL-NVIDIA (Single/Multi) 테스트	• https://catalog.ngc.nvidia.com/orgs/nvidia/containers/hpc-benchmarks
Nemo Framework(Single/Multi) 테스트 Model Customization/Inference	• https://docs.nvidia.com/nemo-framework/index.html

Project Helix 데모센터 고객 제안

Model Customization 및 Inferencing 과정을 Project Helix 데모센터에서 경험할 수 있는 Pilot 환경을 제공



생성형 AI를 위한 델 과 엔비디아의 검증설계를 통해, 안전하고 확장가능한 모듈식 AI 플랫폼을 구축하고 기업맞춤형 AI model을 생성 및 배포해 보세요

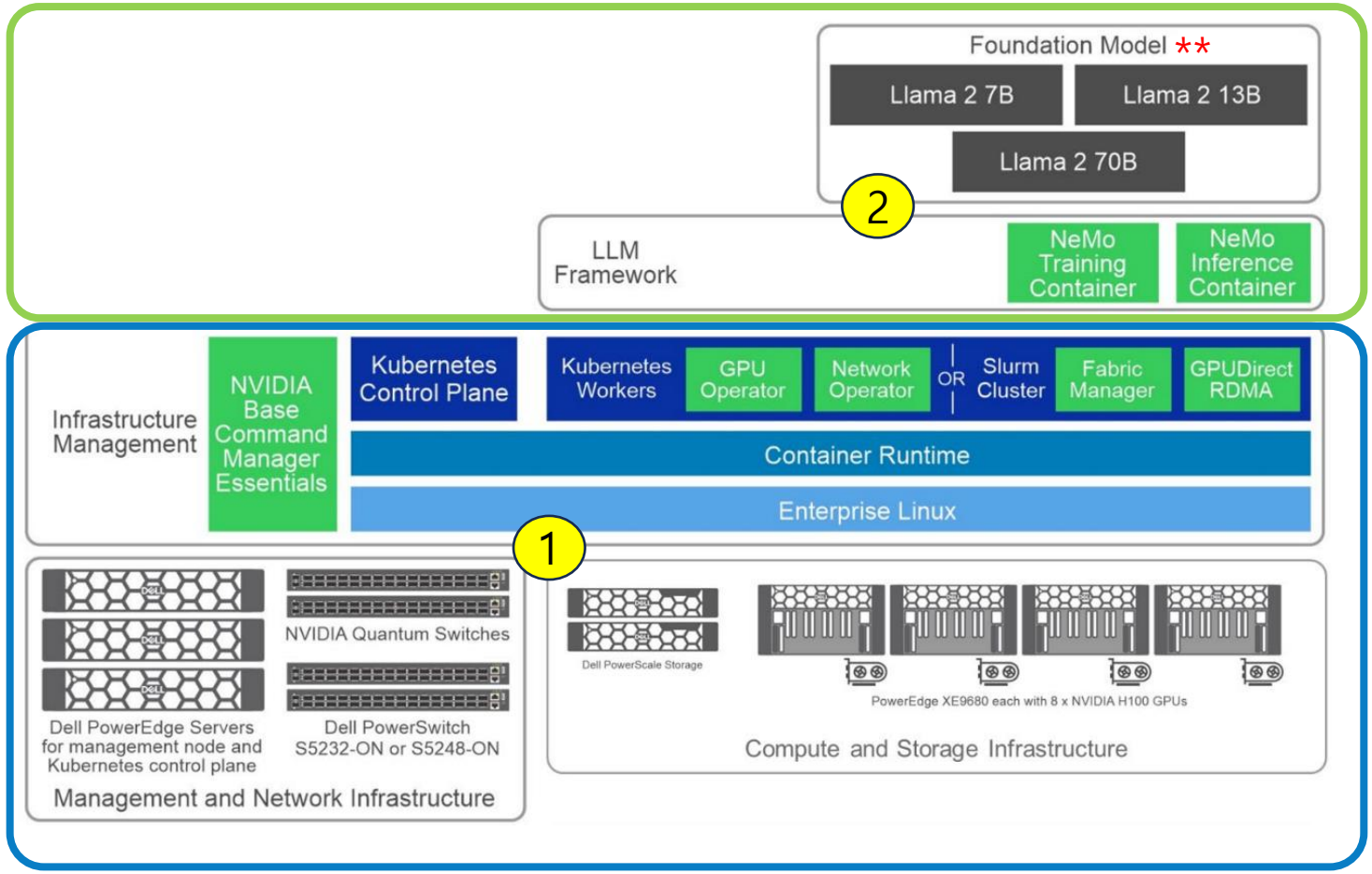
"고객 데이터를 이용한 모델 Customization 및 Inferencing과정을 Project Helix 데모센터를 통해 직접 경험하고 테스트해 보세요. 성공적인 GenAI 전략을 개발하는 데 도움이 됩니다"

Dell의 글로벌 AI 전문가 및 국내 Dell AI 전문 파트너사들이, 여러분의 AI Journey를 같이 합니다.

Customer GenAI Journey

Strategize → Implement → Adopt → Scale

Project Helix 데모센터 Reference Design



1

쉽고 빠르게 최적의 AI 인프라를 구성하고,
관리 및 모니터링

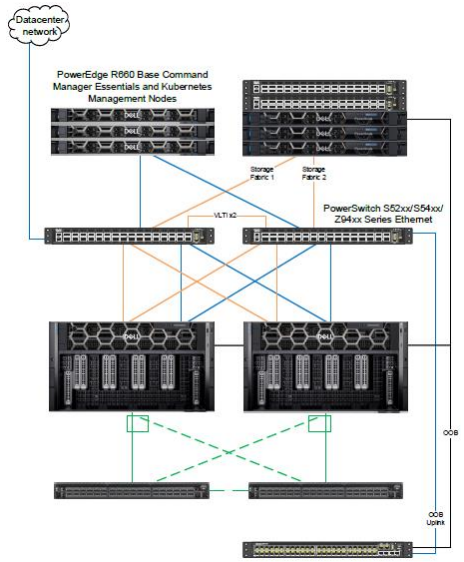
2

LLM Framework를 이용하여 구축된 AI 인프라에서
쉽고 빠르게 모델을 Customization 하고 Inferencing

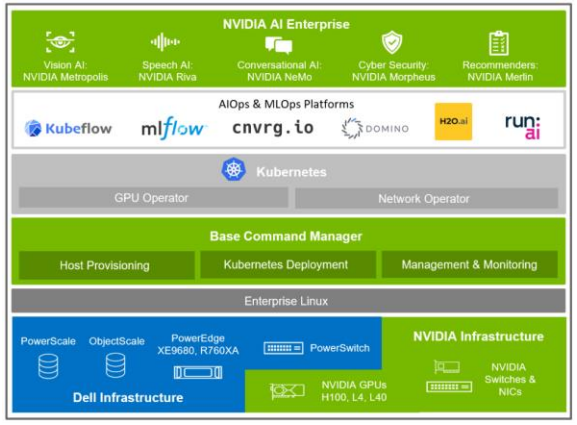
데모센터 구성도

Daol TS MAYMUST

DELL GEN AI in the Enterprise

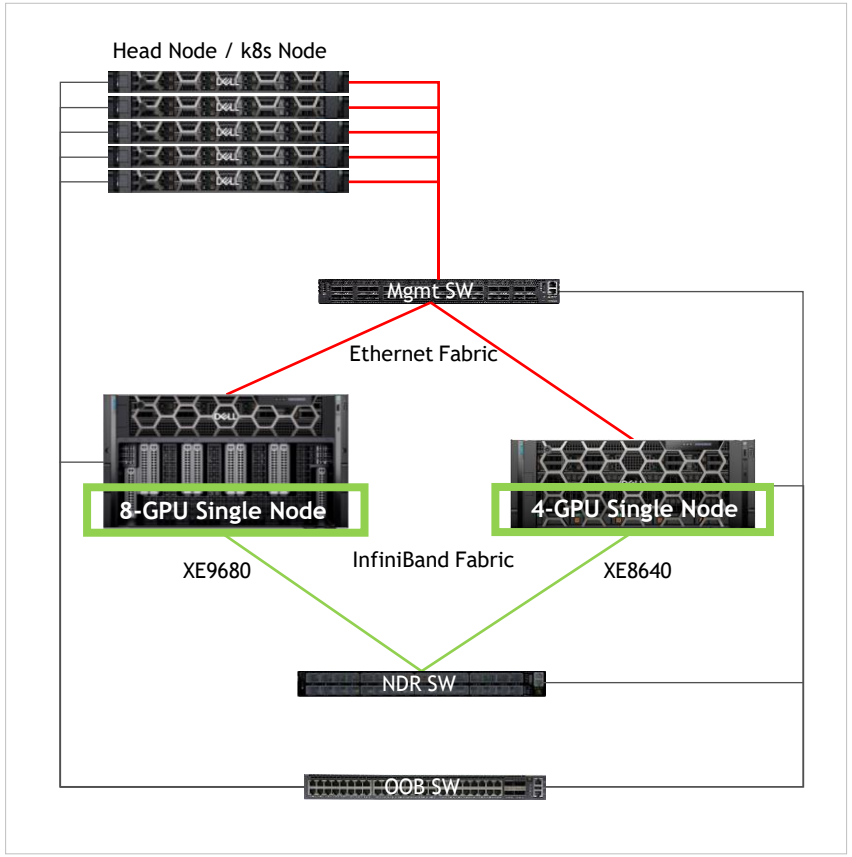


DELL Solution architecture and software stack



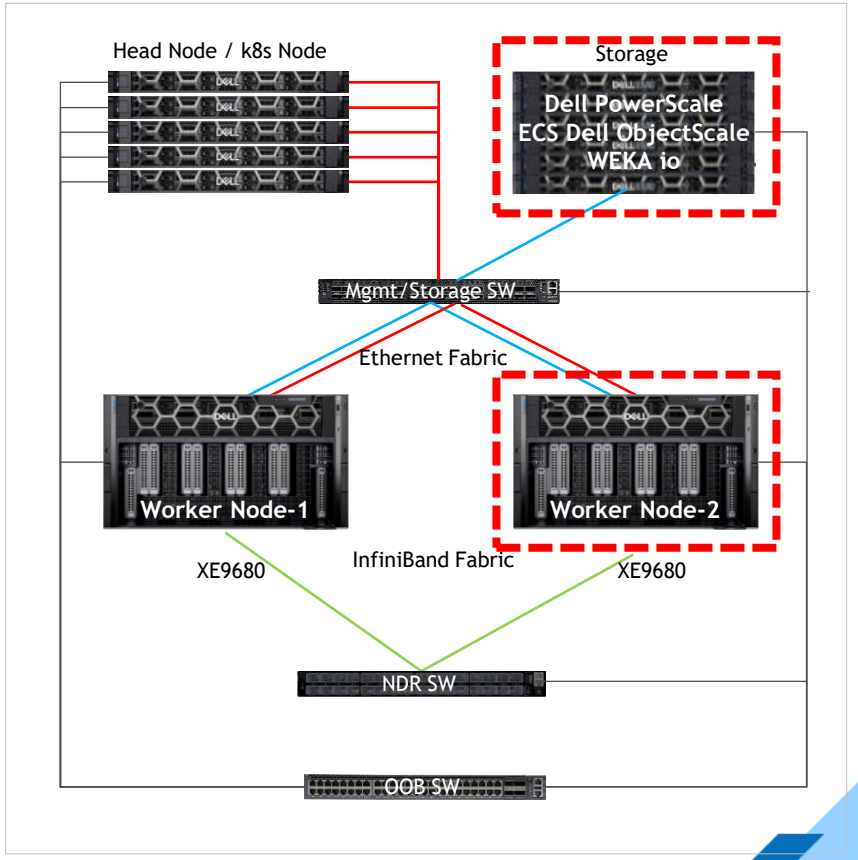
Compute IB NDR(400Gb)
Storage ETH(100Gb)
Mgmt ETH(100GbE)
OOB ETH(1GbE)
ADD-ON

- Single Node -



⟨AS-IS⟩

- Multi Node -

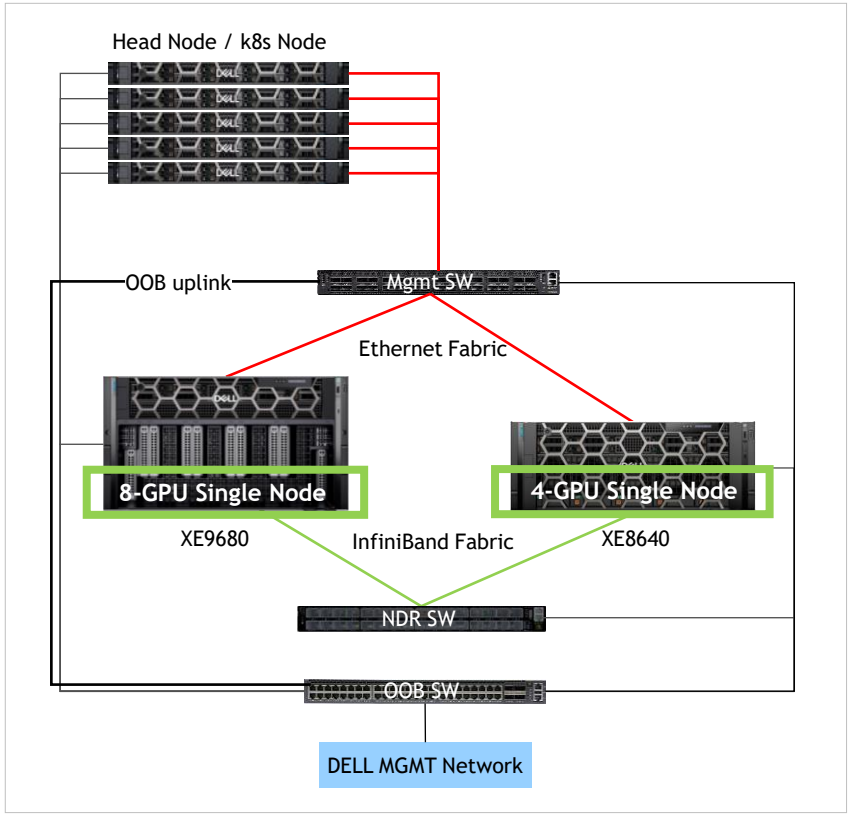


⟨TO-BE⟩

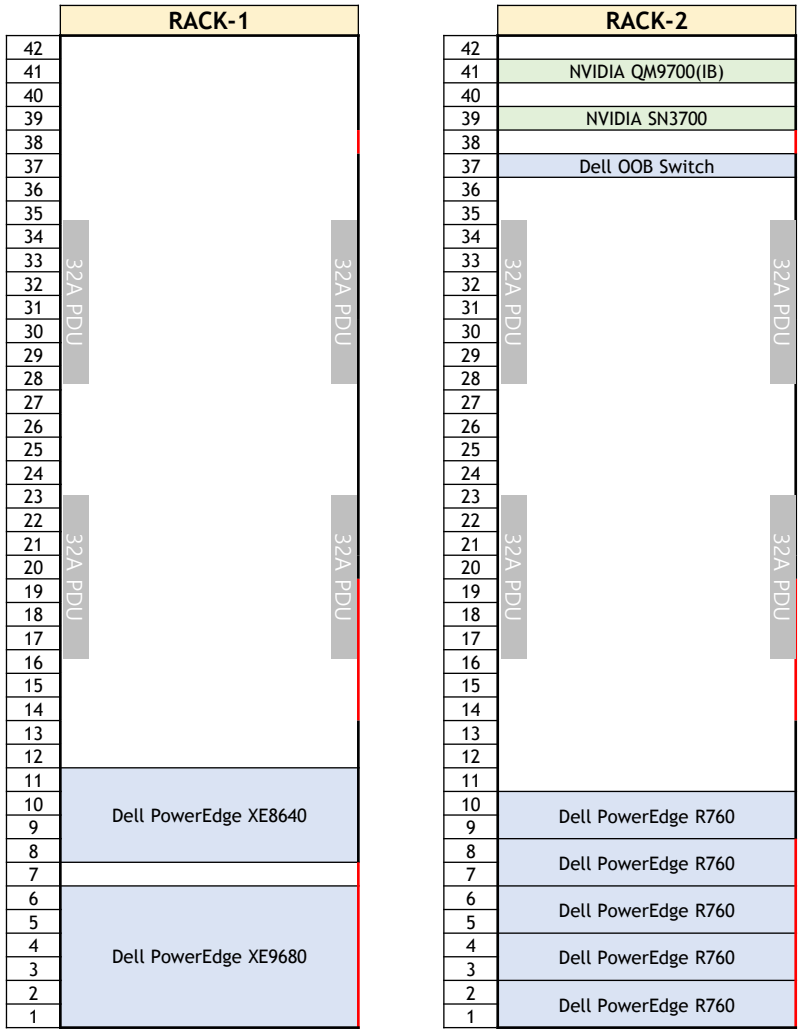
데모센터 구성도

Compute IB NDR(400Gb)
Storage ETH(100Gb)
Mgmt ETH(100GbE)
OOB ETH(1GbE)
ADD-ON

– Single Node –



〈AS-IS〉



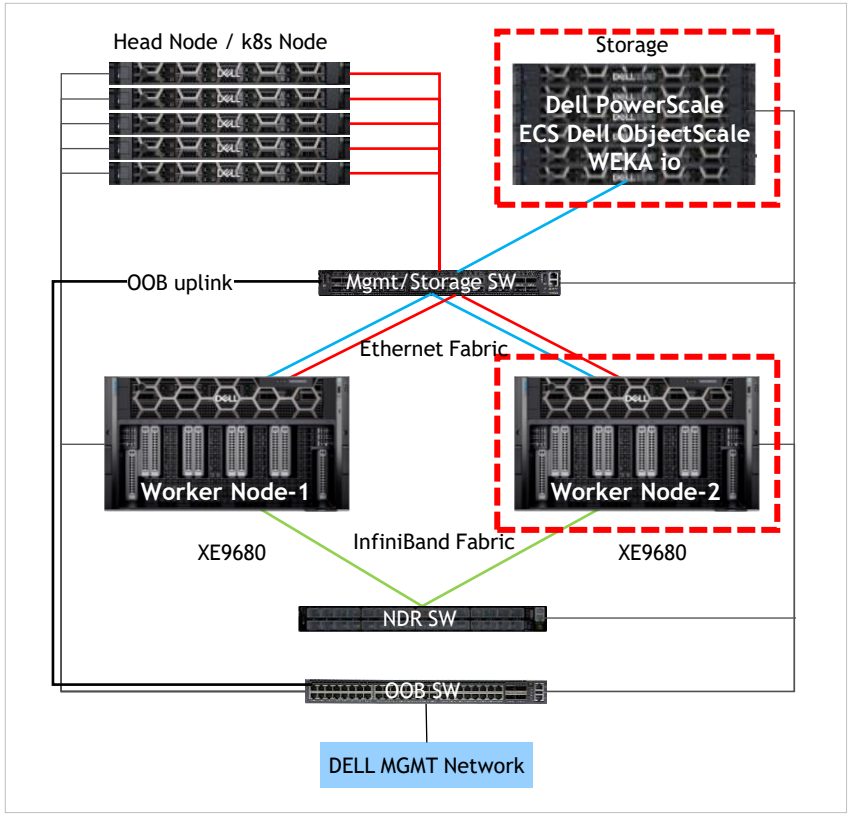
RACK-1	수량	RU	소비전력(W)
Dell PowerEdge XE9680	1	6	12,500
Dell PowerEdge XE8640	1	4	8,400
SUM			20,900

RACK-2	수량	RU	소비전력(W)
NVIDIA QM9700	1	1	1,720
NVIDIA SN3700	1	1	250
Dell OOB Switch	1	1	100
Dell PowerEdge R760	5	10	12,000
SUM			14,070

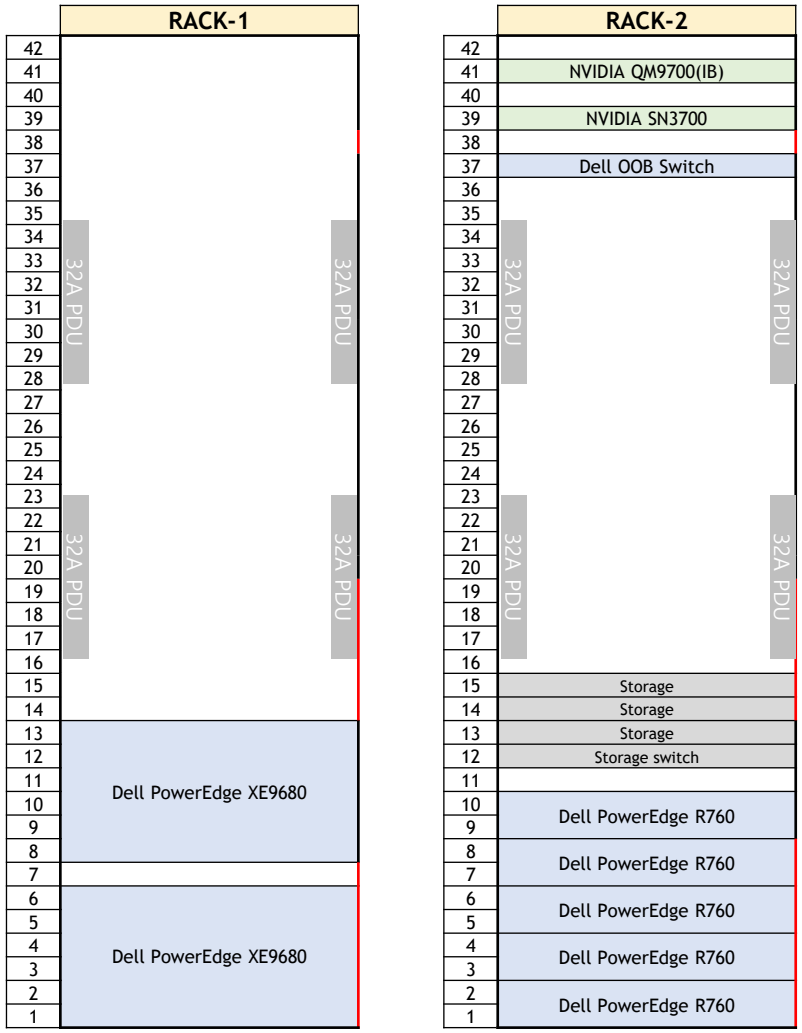
데모센터 구성도

Compute IB NDR(400Gb)
Storage ETH(100Gb)
Mgmt ETH(100GbE)
OOB ETH(1GbE)
ADD-ON

- Multi Node -



〈TO-BE〉



RACK-1	수량	RU	소비전력(W)
Dell PowerEdge XE9680	1	6	12,500
Dell PowerEdge XE8640	1	4	8,400
SUM			20,900

RACK-2	수량	RU	소비전력(W)
NVIDIA QM9700	1	1	1,720
NVIDIA SN3700	1	1	250
Dell OOB Switch	1	1	100
Dell PowerEdge R760	5	10	12,000
Storage	6	6	
Storage switch	1	1	
SUM			14,070

데모 품목

Hardware Parts

구분	Parts	QTY (8GPU, 4GPU)	QTY (8GPU, 8GPU)	비고
Worker Node	PowerEdge XE9680	1	2	HGX H100 8-GPU
	PowerEdge XE8640	1		HGX H100 4-GPU
Head Node / K8s Node	PowerEdge R760	5	5	
InfiniBand Switch	QM9700	1	1	
InfiniBand NDR Transceiver	MMA4Z00-NS	3	8	Switch side
	MMA4Z00-NS400	6	16	GPU node side
InfiniBand MPO Cable	MFP7E10-N010	6	16	
Mgmt Ethernet Switch	SN3700	1	1	
Mgmt Ethernet Switch - Uplink QSA	QSA	1	1	
Mgmt Ethernet Switch - Uplink GBIC	1GbE-T SFP	1	1	
OOB Switch	OOO	1	1	DELL Switch
Ethernet Cable	MFA1A00-C003	5	5	
	MFA1A00-C010	4	4	
NIC	ConnectX-6 DX	2	2	
UTP cable	UTP CAT6 Patch cord	11	11	OOB Switch 연결용

감사합니다.

문의: sales@maymust.com

MAYMUST