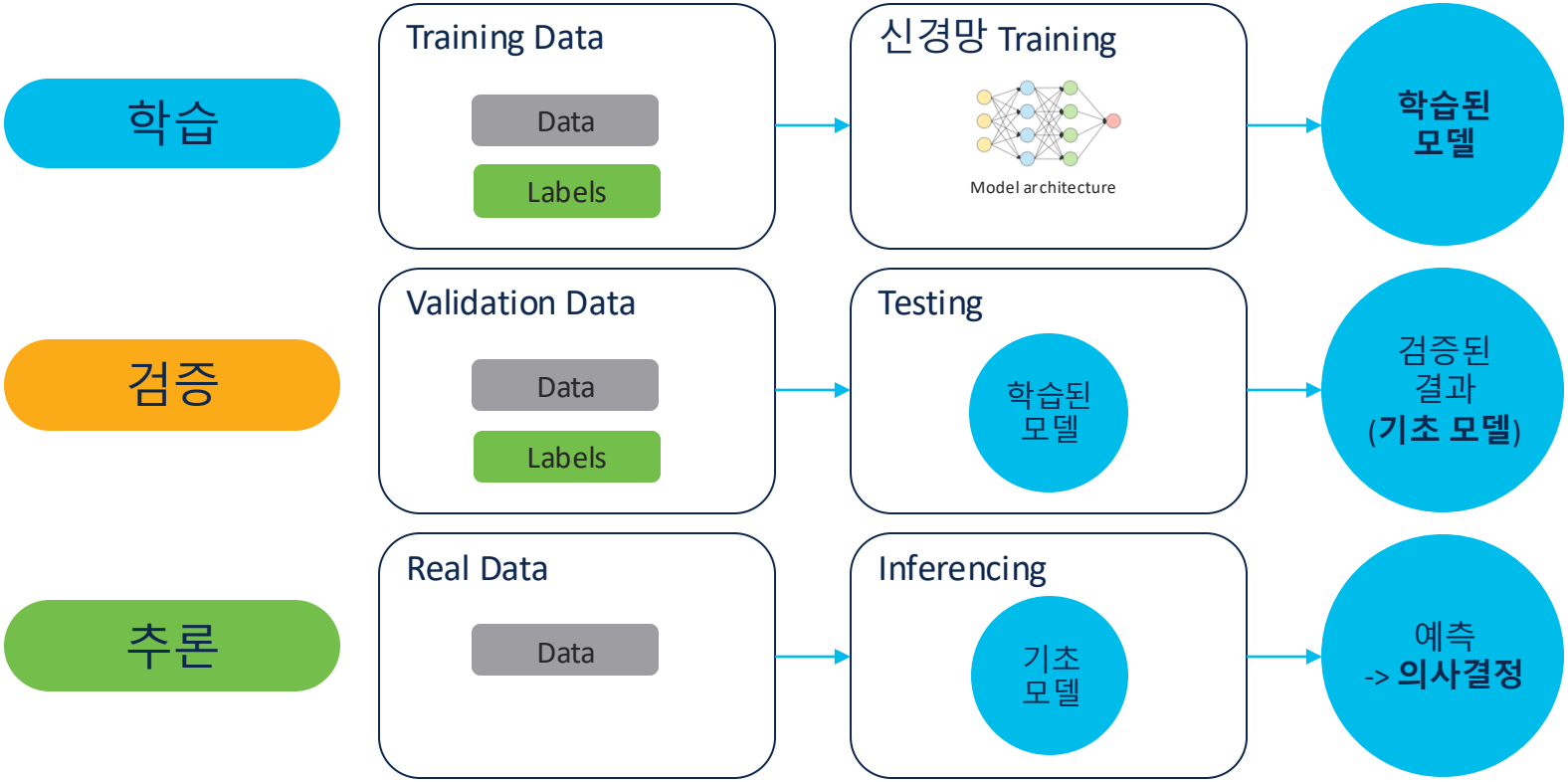
AI 혁신을 위한 최적의 컴퓨팅 솔루션

박한진 프로, 시스코코리아
Cloud and AI Infrastructure Team

Machine Learning Overview



RAG(Retrieval-Augmented Generation-검색 증강 생성)

– LLM 단점을 보완하는 기술

LLM 장점

방대한 지식 보유

언어 이해 능력

다양한 업무능력

편리한 인터페이스

다양한 분야에 사용

LLM 단점

편향성 문제

사실 관계 오류

맥락 이해의 한계

일관성 문제

윤리적 문제

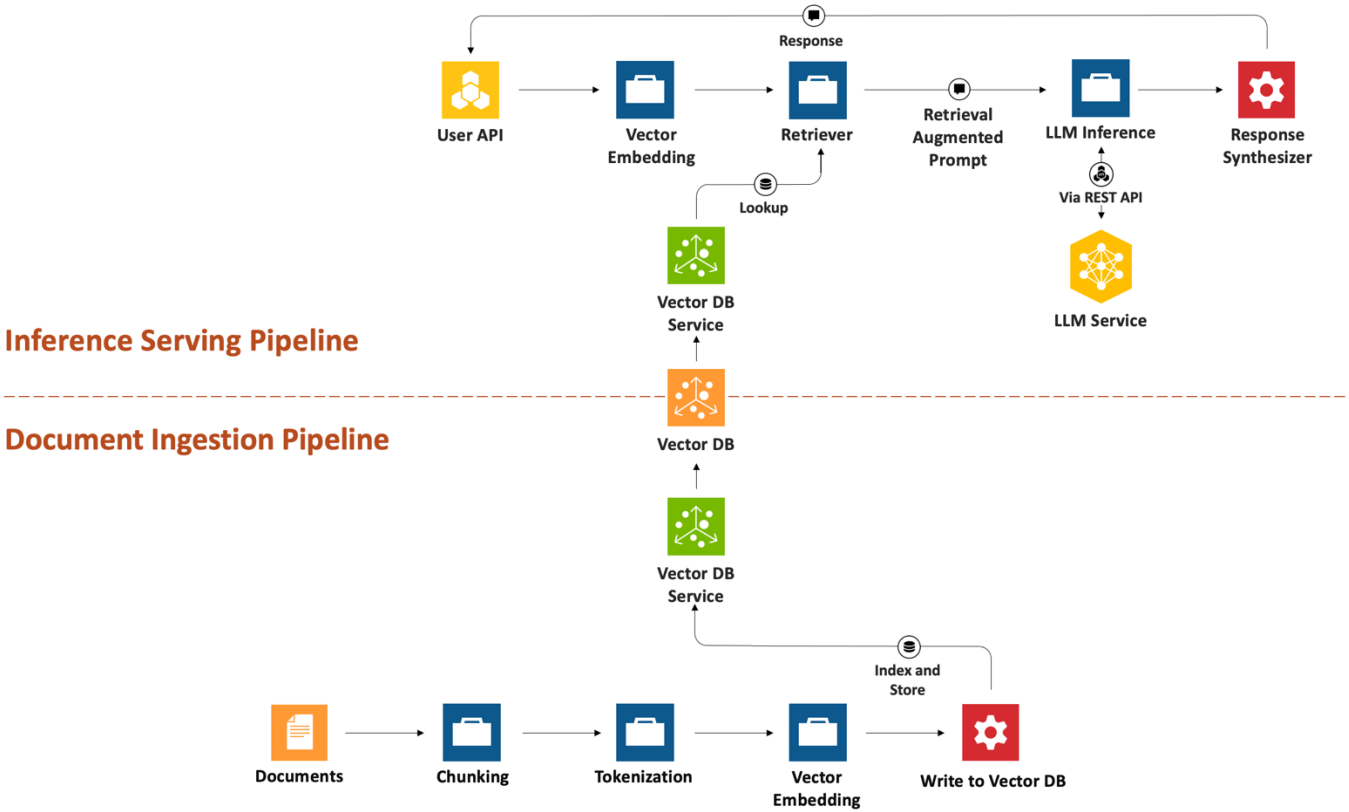
‘사실 관계 오류 가능성’과 ‘맥락 이해의 한계’를 개선하는 데 초점을 맞춤

신뢰성 있는 외부 지식 활용
주어진 질의에 대한 관련 정보를 대규모의 구조화된 지식 베이스(예: 위키피디아)를 모델에 연결 하여 검색 및 추출

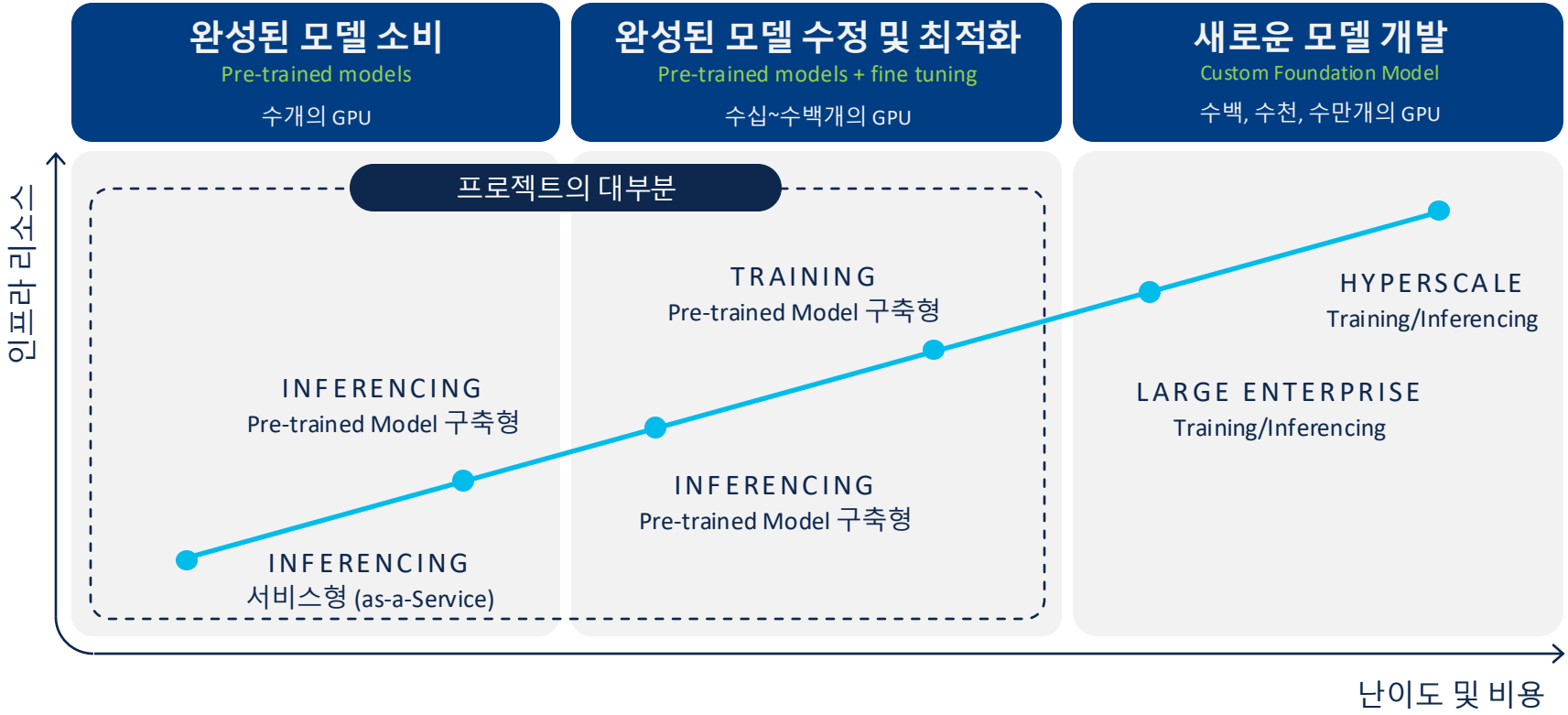
증거 기반 생성 및 출처 제시
검색된 지식 정보를 증거로 활용
생성된 답변의 출처를 명시함

맥락 이해력 향상
외부 지식을 통해 질의에 대한 배경 지식과 맥락 정보를 파악
단순한 패턴 매칭이 아닌 추론 능력을 바탕으로 답변 생성

RAG 아키텍처 – 학습된 모델을 나의 업무에 적용시 반드시 고려



LLM을 고려하는 고객의 스펙트럼



산업별 AI/생성형 AI 활용 사례 예시




지식기반 코파일럿
AI 비서

컨텐츠와 코드 생성
텍스트 | 이미지 | 비디오 | 코드

리포팅과 데이터 분석
텍스트 요약 | 시각화 생성

언어 번역
다국어 실시간 커뮤니케이션

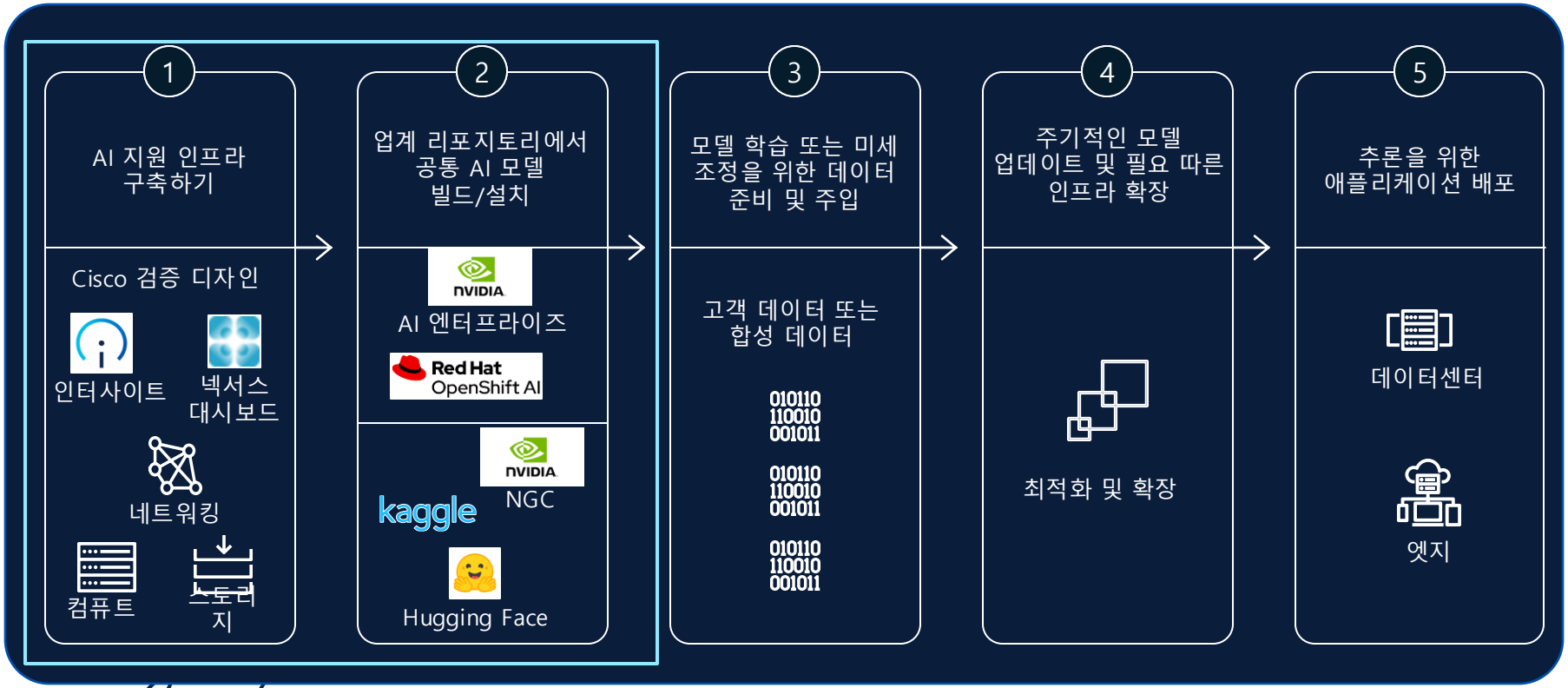
가상 에이전트 & 챗봇
도메인별 특화된 챗봇

탐지 & 예측
예측 | 이상 징후 | 인사이트




모델 구축 및 학습 | 모델 최적화 및 미세 조정 | 모델 활용 및 추론

AI 개발 프로젝트의 전형적인 워크 플로우 예시



SLM = 비용 + 전문성 + 온디바이스 AI

도메인 특화에 사용되는 SLM(Small Language Model or Specialized Language Model)

헬스케어&생명과학

진료 기록 요약
의료 데이터 기반 응답
개인 의료 정보 관리

금융 서비스

사기 탐지
위험성 평가
거래

공공 부문

스마트 시티
안전
서비스 개선

유통&소매

개인화
인벤토리 최적화
매출 예측

제조

예측 가능한 유지 관리
품질 관리
수요 예측

에너지

배포 최적화
결함 예측
수요 예측

농업

수율 최적화
자동 관개
해충 예측 및 예방

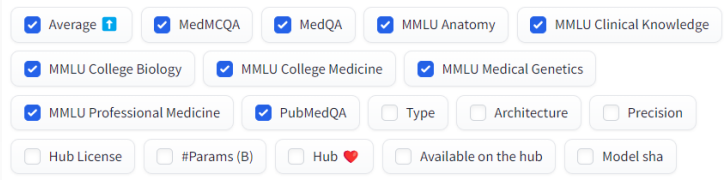
수송

경로 최적화
자율 주행 차량
예측 가능한 유지 관리

IT

장애 분석
코드 작성 및 리뷰
유지관리 방안 제시

Gen AI(Generative AI)
의료 도메인 특화 SLM



usmle

T	Model	Average	MedMCQA	MedQA	MMLU Anatomy	MMLU Clinical
◆	jiviai/medX_v1	91.65	77.38	85.94	95.56	97.74
◆	bongbongs/NewMes-v10.2.1	90.79	73.92	80.83	97.04	97.74
◆	jiviai/medX_v0	90.69	75.07	81.54	97.78	96.23
◆	bongbongs/NewMes-v8.3	90.57	74.56	80.83	97.04	96.98
◆	bongbongs/NewMes-v8.3	90.53	74.35	81.15	97.04	96.98
◆	bongbongs/NewMes-v10	90.15	72.77	80.91	94.81	96.23
◆	bongbongs/NewMes-v10.2	90.08	74.59	81.23	95.56	96.98
◆	bongbongs/NewMes-v10.1	90.07	74.47	79.65	97.04	94.72
◆	bongbongs/NewMes-v8.4	90.04	75.81	81.46	93.33	96.98
◆	ProbeMedicalYonseiMAILab/medllama3-v20	90.01	75.4	81.07	91.85	95.85
◆	ProbeMedicalYonseiMAILab/medllama3-v20	89.94	75.19	81.38	91.85	95.47
◆	bongbongs/NewMes-v9	89.68	73.85	80.44	96.3	96.23

Gen AI(Generative AI)
금융 도메인 특화 SLM

Model Information | Information Extraction (IE) | Textual Analysis (TA)
Question Answering (QA) | Text Generation (TG) | Risk Management (RM)
Forecasting (FO) | Decision-Making (DM) | Spanish | Other

T	Model	Average	Average IE	Average TA	Average QA	Average TG	Average RM
●	GPT4	39.2	35	64.4	50.7	10	51.7
●	LLaMA3.1-70B	36.2	15.7	63.6	14.7	9	0
●	Qwen2-72B	34.7	12.6	59.5	0.3	11	0
●	Xuanyuan-70B	34.4	9.3	61.4	0.7	12.5	0
●	LLaMA3.1-8B	34.3	15.6	56.2	1.3	10	0
●	Gemini	32.4	22.1	58.4	20.3	19.5	51.8
●	ChatGPT	29.2	26.4	59	39.3	8.5	45.6
●	meta-llama/Llama-2-70b	25.8	10.6	59.9	10.7	12.5	50

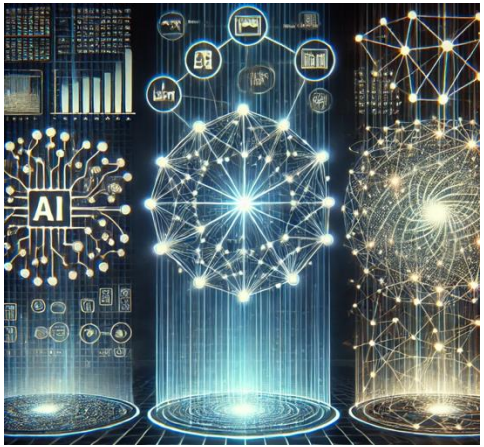
Model	Average	Average IE	Average TA	Average QA	Average TG	Average RM	Average FO	Average DM
GPT4	39.2	35	64.4	50.7	10	51.7	54.3	75.2
Gemini	32.4	22.1	58.4	20.3	19.5	51.8	53.7	67.2
LLaMA3.1-70B	36.2	15.7	63.6	14.7	9	0	46	49.3

AI프로젝트 시작 시 고려사항 - 완성도

데이터 품질 및 양

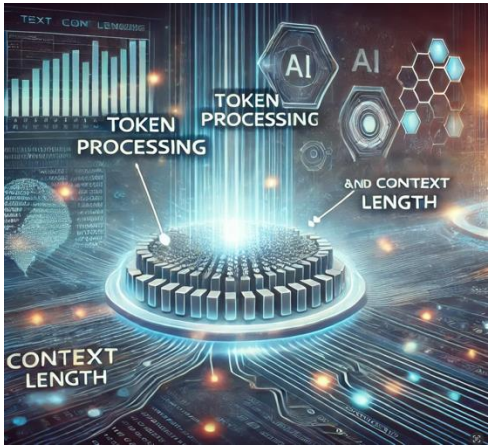
데이터의 다양성 ,정확성
데이터 정제
적절한 데이터 크기

모델의 종류, 파라미터 수



적절한 모델 및
파라미터 선택
자원과 비용을 고려

토큰 처리 및 문맥 길이



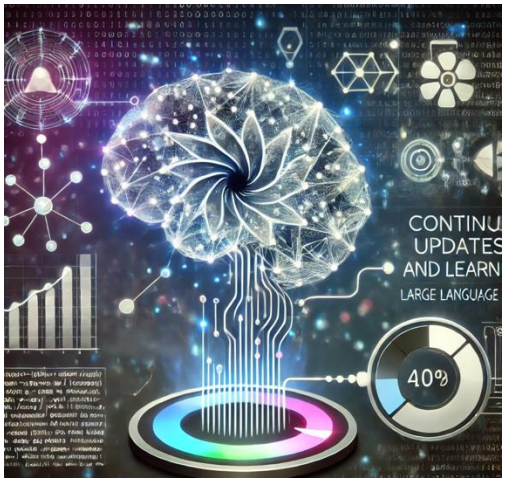
토큰(Token)의 개수
긴 문맥 유지
문서 분할 기법

AI프로젝트 시작 시 고려사항 - 비용

추론 속도 및 비용

연산 속도
컴퓨팅 비용
양자화(Quantization)

지속적인 업데이트 및 학습



모델 업데이트 및 튜닝
RAG 연동
RHEL – 사람의 피드백

활용 환경 및 배포 고려

Onprem vs Cloud
Edge AI
API 호출 비용

그리고 AI 를 위한 인프라

컨테이너 플랫폼

AI 전문 툴

운영 및 자동화

SERVER & GPU

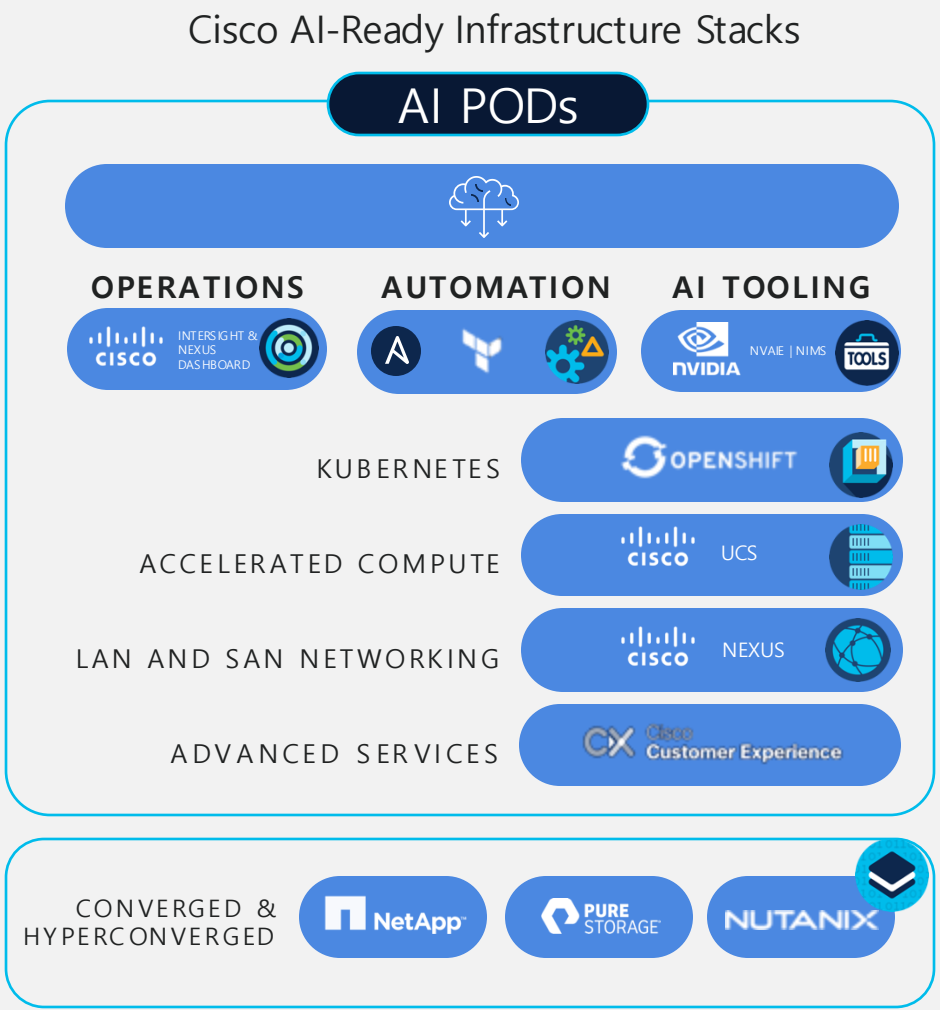
스토리지

네트워크

Cisco AI PODs

번들 형태로 Full Stack 빠르게 오더

- Pre-Designed 된 Bundles 제품
- 빠른 배포를 위한 Framework
- 인프라 운영관리 솔루션 제공
- 검증된 Use Cases 제공
- Adoption Services 포함 (초기 구성)
- Full Stack Support
- AI 여정을 시작하는 고객을 위한 솔루션



cisco Connect

AI-POD 를 통한 고객의 기대 가치

추론을 위해 고려되는 다양한 모델의 Sizing 및 성능치 제공 -> 도입 시 사용할 모델에 대한 기대 성능 확인

- Stable Diffusion
- Openjourney
- Dreamlike Diffusion 1.0
- Hotshot-XL
- Llama 2
- Llama 2 Inferencing with Pytorch
- Nemotron-3 8B Models
- GPT-2B
- FLAN-T5
- Mistral 7B
- BLOOM
- GALACTICA
- Falcon-40B
- Defog SQLCoder
- Code Llama
- GPT-NeoX-20B
- MPT-30B
- OPT : Open Pre-trained Transformer Language Models

Llama 2 – 7B
Input tokens Length: 128 and output Tokens Length: 20

Batch Size	GPUs	Average Latency (ms)	Average Throughput (sentences/s)
1	1	241.1	4.1
2	1	249.9	8.0
4	1	280.2	14.3
8	1	336.4	23.8
1	2	197.1	5.1
2	2	204.1	9.8
4	2	230.2	17.4
8	2	312.6	25.5

Figure 120. Latency of Llama 2-7B-Chat with input tokens: 128 and output tokens: 20 with 2 GPU

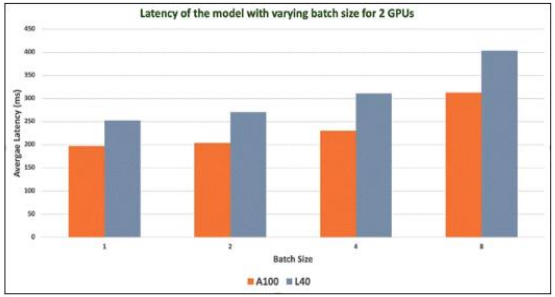
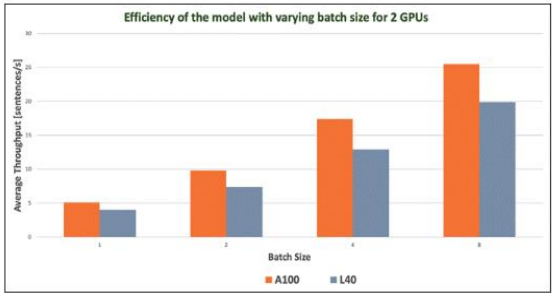


Figure 121. Throughput of Llama 2-7B-Chat with input tokens: 128 and output tokens: 20 with 2 GPU



Optimized price to performance ratio with FLASHSTACK AI

AI-POD 를 위한 Sizer 제공

운영과 유사한워크로드 및 사이징

Workload Inputs

Field	Value
Operation	Inferencing
Model	Llama-3.1-8B
Target Precision	FP16 / BFLOAT16
Input Sequence Length	512
Output Sequence Length	1024
Concurrent Users (Batch Size)	2000
Max GPU Memory Utilization (%)	100

Compute Summary

Type	Option 1	Option 2
Server Count	7	13
Server Type	C240 M7 SFF	C240 M7 SFF
CPU Type	6548Y+	6548Y+
Server Memory (GB)	512 (32x16GB)	512 (32x16GB)
GPU Type	H100	L40
Total GPU Count	13	26
Memory (GB) / GPU	80	48
TFLOPs / GPU	756.0	181.05

Performance Metrics (values are per GPU)

Metric	Option 1	Option 2
Time to First Token (TTFT) (ms)	5911.83	11711.19
Time per Output Token (TPOT) (ms)	30.27	30.65
Latency (Response Time) (s)	36.88	43.07
Throughput (Tokens/s)	5087	2511
Batch Size	154	77
Model Memory Size (GB)	16.06	16.06
Activation Memory Size (GB)	62.01	31.00
Total Memory Required (GB)	78.07	47.06
GPU Memory Utilization (%)	97.59	98.05

Cisco Solution Attached Services for Cisco AI PODs

AI PODs 구축 지원 서비스

아키텍처 설계	설치 및 구성	테스트 및 검증	최적화 지원	지식 이전
• 표준화된 AI POD 아키텍처를 활용한 솔루션 설계 • 요구사항에 맞는 적절한 구성요소 선정 (서버, GPU, Openshift 등)	• 관리 솔루션 • 서버 및 GPU • Red Hat Openshift • NVIDIA GenAI 소프트웨어 스택 • 자동화 도구	• 모범사례 및 벤치마크를 준수한 종단간 테스트 • OpenShift, NVIDIA GenAI 소프트웨어 스택, storage연동 검증	• AI POD 인프라 성능 관련 튜닝 및 최적화	• AI POD 인프라 구현에 대한 지식 이전 세션 제공

MGX? HGX? DGX ?


Platform Software
Networking
Switch
InfiniBand Quantum-X (400G)
Ethernet Spectrum-X (400G)
NICs (including DPUs)
Connect-X NIC
BlueField Series DPUs
DGX BasePod Reference Architecture
Storage
Typically provided by Partner e.g., DDN, IBM, NetApp, Vast (Certified Storage)
Compute (AI Servers)
Inter-connects
Systems
NVSwitch
Chip to Chip
NVLink C2C (CPU to GPU)
NVLink (GPU to GPU)
Other Server Components
Semi-Conductors
DPUs
BlueField-3 & 4
CPUs
Grace Series
GPUs
H100, H200, B100 - Training
L4, L40, L40S - Inference
GH200, GH200NVL (CPU+GPU Superchips)
MGX Reference Architecture
HGX Server Platforms
DGX Servers (includes SSD)
Offerings
Offers
Point Products/ Tech

HGX Nvidia Reference Architecture

NVIDIA OVX(**Omniverse eXtended**) Systems :
2-8-9-400 (CPU-GPU-NIC)

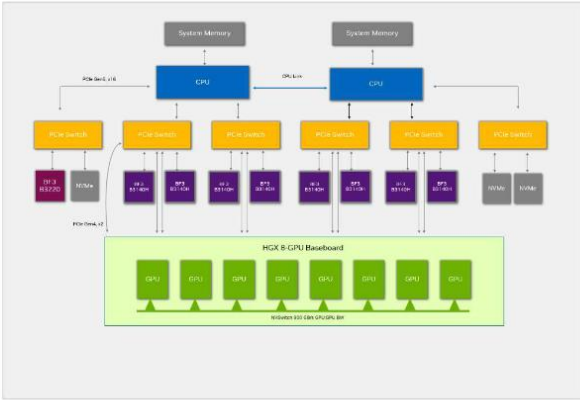
2개의 CPU
8개의 SXM GPU
(예: SXM타입 H100/H200)
9개의 네트워크 어댑터
(예: ConnectX-7, BlueField-3 SuperNIC
으로 구성하는
NVIDIA 인증 스케일아웃 컴퓨팅 노드를
정의합니다.

이 구성은 클러스터에서 최대 4개에서
최대 64개 노드까지 확장할 수 있습니다

NVIDIA HGX H100/H200 Enterprise Reference Architecture

HGX 2-8-9 Configuration

NVIDIA HGX H100/H200 — SERVER	
CPUs	2x 56c Intel Xeon Gold 8480+ 2x 64c AMD EPYC 9554
GPUs	8x NVIDIA H100/H200 SXM
Networking — E/W	8x BlueField-3, B3140H (1x400Gb)
Networking — N/S	1x BlueField-3, B3220 (2x200Gb)
Host Memory	Min 1.5 TB DDR5 ECC (1 DIMM per slot)
Host Boot Drive	1x 1TB NVMe
Host Storage	4x 4TB NVMe



https://docs.nvidia.com/certification-programs/latest/nvidia-certified-systems.html?utm_source=chatgpt.com#pcie-optimized-reference-configuration-2-8-5-200

UCS C885A M8 GPU 서버 (HGX)

모델 트레이닝과 딥 러닝과 같은 고성능 사용에 최적화된 UCS C885A

UCS Accelerated | UCS c885A M8

2 CPUs
AMD Genoa (96 cores)/Turin

NVIDIA HGX with 8 GPUs
NVIDIA H100, H200, B100 with NVLink
AMD MI300 with Infinity Fabric

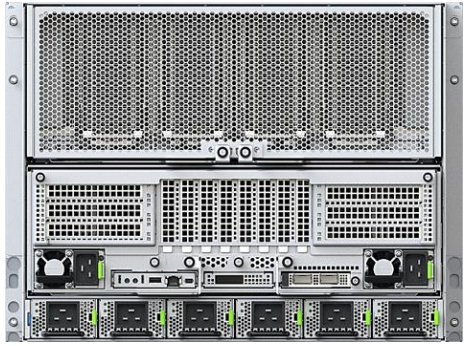
Network
(8) NVIDIA ConnectX-7 BF3 B3140H (E-W)
(1) NVIDIA BF3 B3220, B3240 (N-S)

Power
(6) 3000W and (2) 2400W

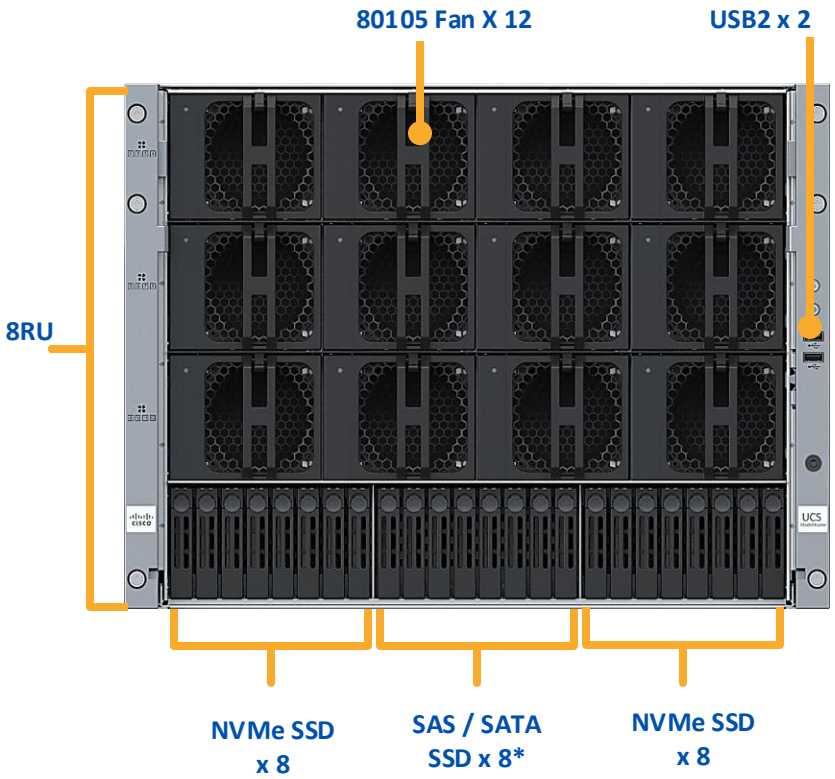
Air Cooling

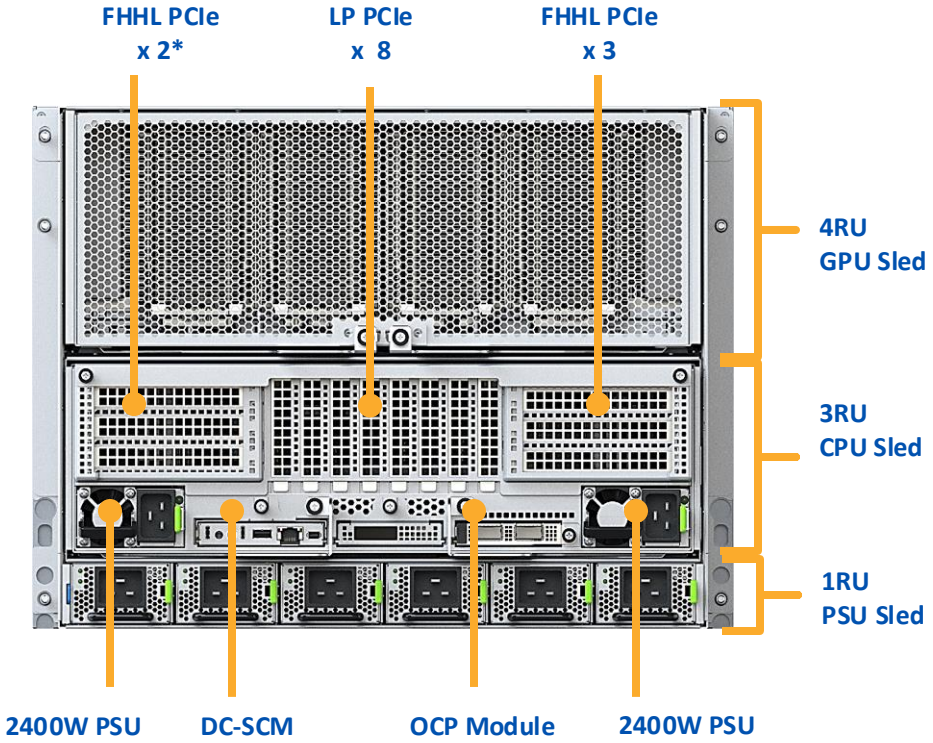
UCS C885A M8 상세규격

Product Specifications	
Form Factor	• HGX 8U 19” EIA Rack
Compute + Memory	• 2 AMD EPYC 4th (Genoa) or 5th (Turin) Gen CPUs • 24 DDR5 RDIMMs • Up to 6,000 MT/S
Storage	• 1 PCIe3 x4 M.2 NVMe (Boot Device) • 16 PCIe5 x4 2.5” U.2 NVMe SSD (Data Cache)
GPU	• 8 H100 700W or 8 H200 700W or 8 B200A 700W • 8 MI300X 750W
Network Cards	• 8 PCIe5 x16 HHHL for E-W NIC ConnectX-7, BF3 B3140H • 5 PCIe5 x16 FHHL for N-S NIC BF3 B3220, B3240 (max 2) • 2 OCP 3.0 SFF
Cooling	• 12 80105 Hot swappable (N+1) fans for system cooling • 4 6056 fans for SSD cooling
Front IO	• 2 USB 2.0, 1 ID BTN, 1 Power Button
Rear IO	• 1 USB 3.0 A, 1 USB 3.0 C, mDP, 1 ID BTN, 1 Power Button, 1 USB 2.0 C (for debugging), 1 RJ45 (mgmt.)
Power Supply	• Up to 6 54V 3kW (N+2) and 2 12V 2.7kW MCRPS/CRPS, N+1 redundancy

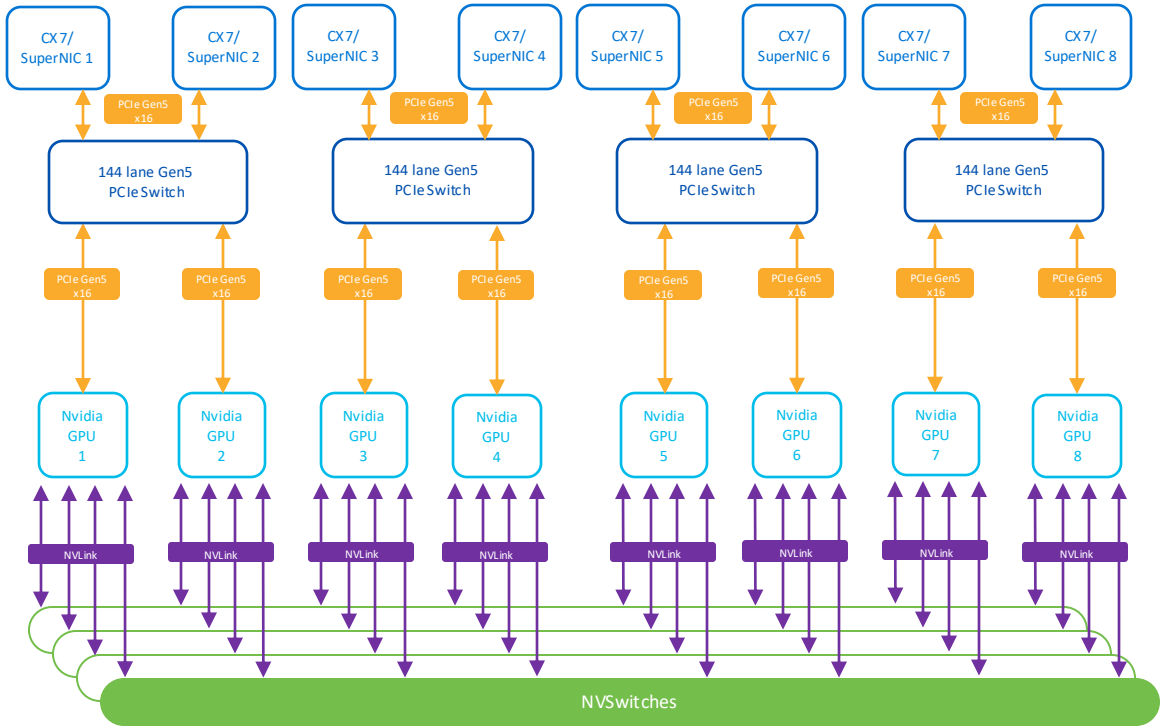


C885A M8: 패널 I/O 위치



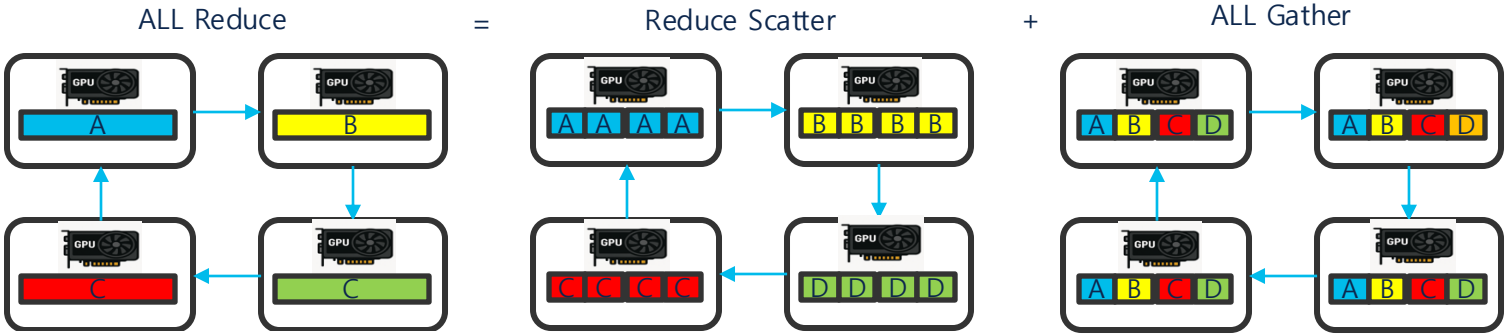


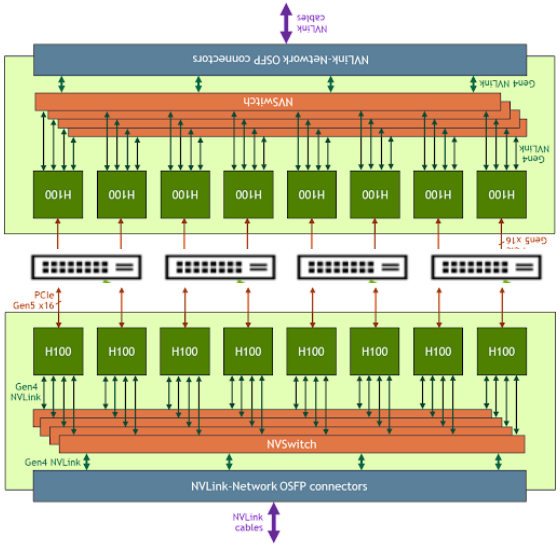
UCS C885A – Nvidia GPU 연결 내부 네트워크



- 8개의 Nvidia H100/H200 SXM5 Tensor Core GPU
- 각 H100/H200 GPU에는 여러 개의 NVLink 포트가 있으며 4개의 NVSwitch에 모두 연결됩니다.
- 8개의 모든 GPU를 연결하는 4개의 NVSwitch
- 동일한 노드에 있는 모든 GPU 쌍 간의 NVLink 양방향 속도는 900GB/s
- 각 H100/H200 GPU에는 노드 간 GPU 간 연결을 위해 PCIe Gen5 x16을 통해 연결된 전용 NIC/SuperNIC도 있습니다.

WHY NIC is key? WHY400G NIC?





항목	**NVLink 4.0**	**400G NIC**
단일 링크 단방향	50 GB/s (1 link)	50 GB/s (400 Gbps ÷ 8)
단일 링크 양방향	**100 GB/s**	**100 GB/s**
H100 기준 전체	18 links → **900 GB/s (양방향 총합)**	일반적으로 1 NIC → **100 GB/s**
통신 대상	GPU 간 고속 메모리 직접 접근 (intra-node)	서버 간 통신 (inter-node)
지연시간	수십 ns	수 µs (TCP/IP 기준)

MGX Nvidia Reference Architecture

NVIDIA OVX(**Omniverse eXtended**) Systems :
2-8-5-200 (CPU-GPU-NIC)

2개의 CPU
8개의 PCIe GPU
(예: L40S, H100 NVL)
5개의 네트워크 어댑터
(예: ConnectX-7, BlueField-3 SuperNIC)
으로 구성하는
NVIDIA 인증 스케일아웃 컴퓨팅 노드를
정의합니다.

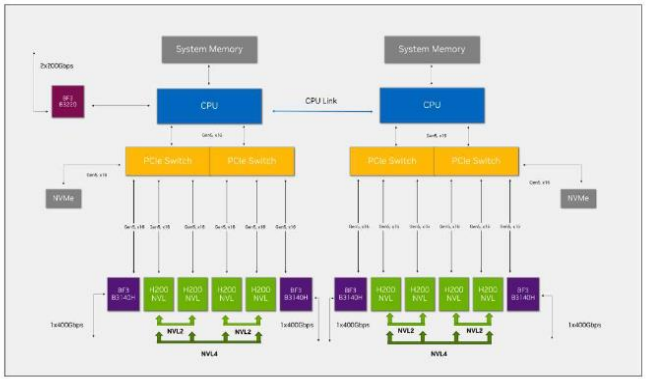
이 구성은 클러스터에서 최대 4개에서
최대 64개 노드까지 확장할 수 있습니다

NVIDIA H200 NVL Enterprise Reference Architecture

PCIe-Optimized 2-8-5 Configuration

H200 NVL— SERVER	
CPUs	2x 56c Intel Xeon Gold 8480+ 2x 64c AMD EPYC 9554
GPUs	8x NVIDIA H200 NVL
Networking — E/W	4x BlueField-3, B3140H (1x400Gb)
Networking — N/S	1x BlueField-3, B3220 (2x200Gb)
Host Memory	Min 1,024GB DDR5 ECC (1 DIMM per slot)
Host Boot Drive	1x 1TB NVMe
Host Storage	Min. 2x 4TB NVMe (one per socket)

NVIDIA Certified



https://docs.nvidia.com/certification-programs/latest/nvidia-certified-systems.html?utm_source=chatgpt.com#pcie-optimized-reference-configuration-2-8-5-200

UCS C845A M8 GPU 서버 (MGX)

Cisco UCS® C845A M8 Rack Server

NVIDIA MGX 플랫폼

2/4/8 NVIDIA H100 NVL/H200 NVL/L40S GPU

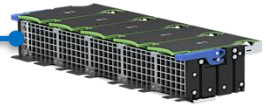
2 AMD 5세대 EPYC 프로세서

H100*8+B3220+B3140H*4

L40S*8+B3220+CX7*4

UCS C845A M8 GPU 서버 (MGX)

Upper Cooling/Hot Swap
- 5x 8080 Fan



M.2

Host Processor Module

Storage
- 20x E1.S

Lower Cooling/Hot Swap
- 5x 8080 Fan

PCIe E-W Option
- 4x CX7 (1x400G)
- 4x BF3 B3140H

PCIe N-S Option
- 1x BF3 B3220 (2x200G)
- CX7 (2x200G)
- 1x BF3 B3240 (2x400G) (Double Wide)

GPU Cards
- 8x H100 NVL
- 8x H200 NVL
- 8x L40s
- 8x MI210

Power Supply

OCP 3.0 NIC (RJ45)
- Intel X710-DA2 (2x10GbE)

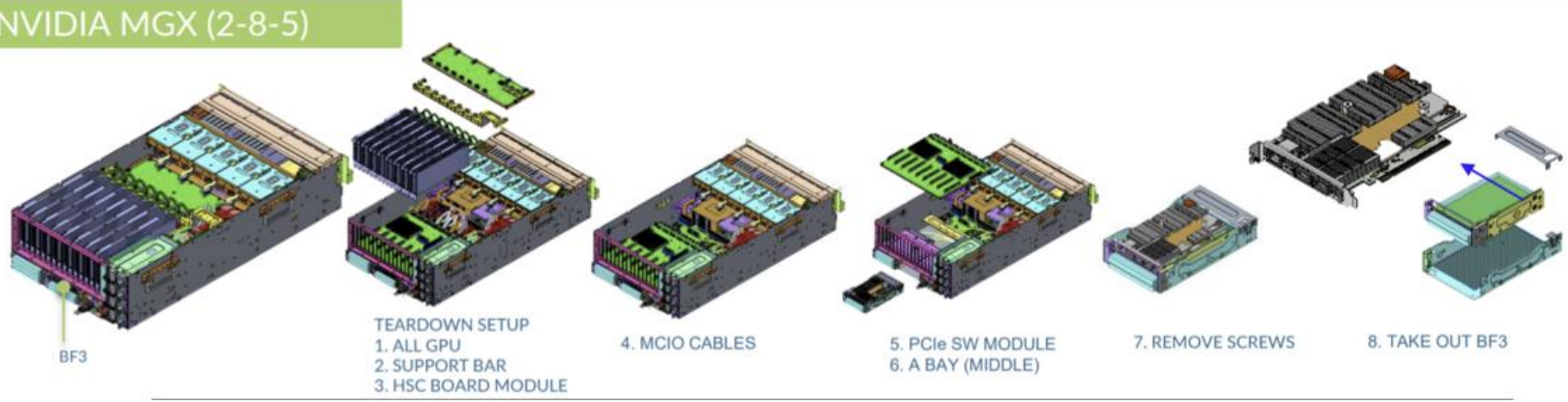
DC Secure Control Module (BMC)

4RU Chassis
- D 800 mm
- W 438 mm
- H 175 mm
- 60kg/132lb

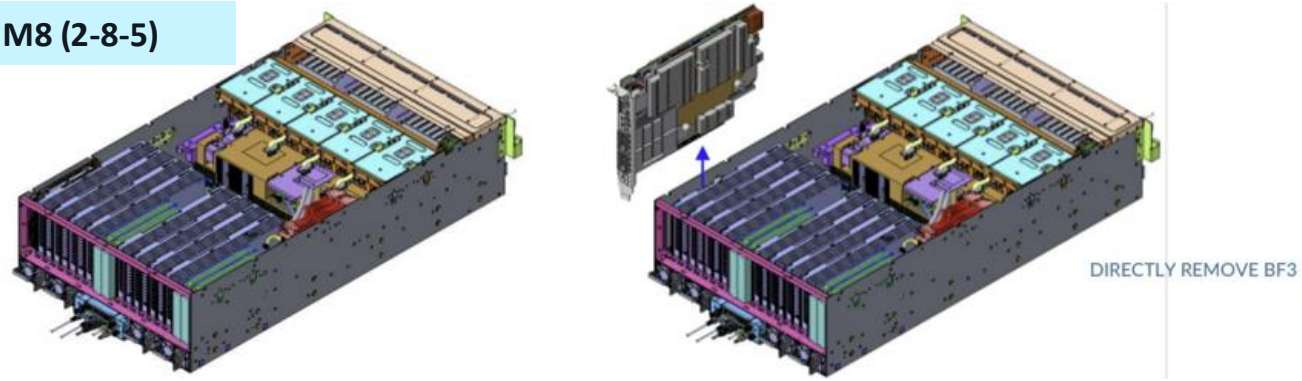
The UCS C845A M8 – A Better Overall Design

PCIe Card Serviceability Example

NVIDIA MGX (2-8-5)

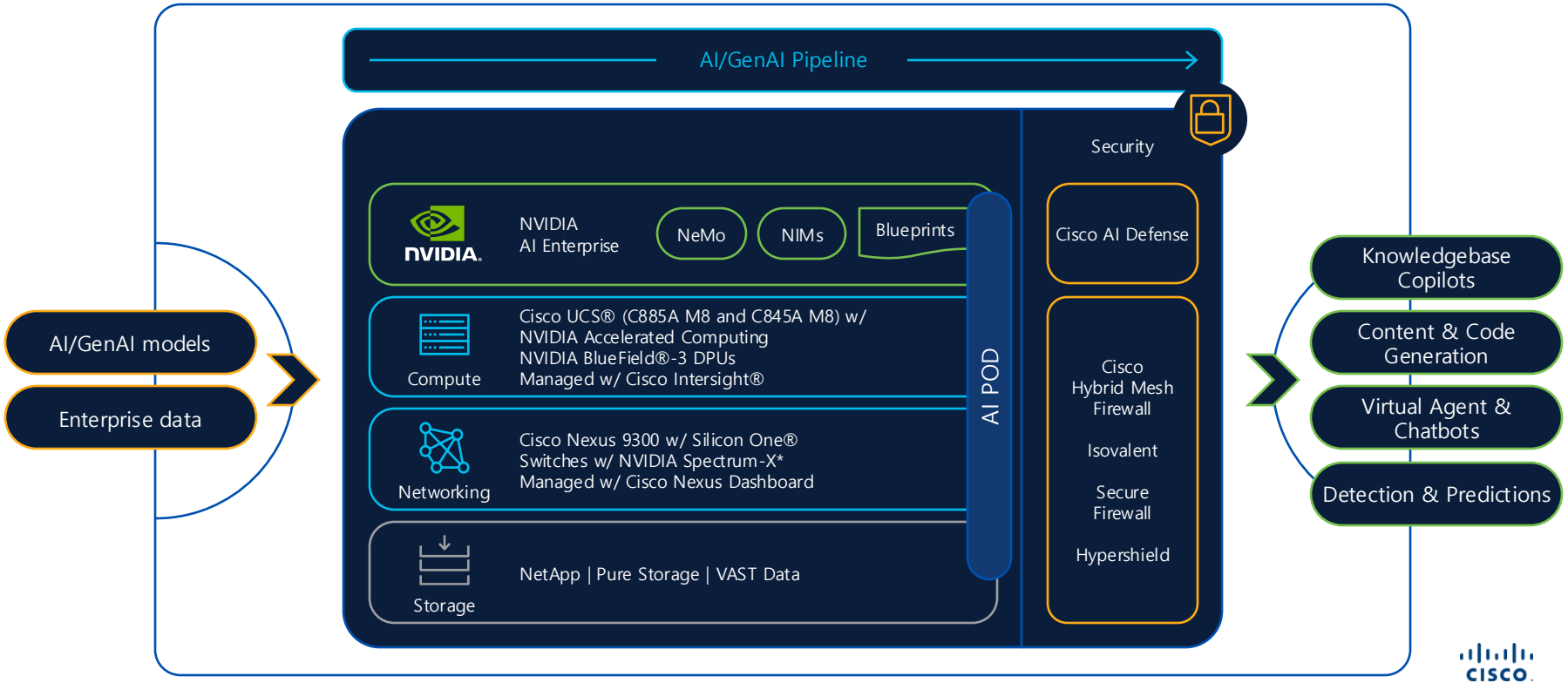


Cisco UCS C845A M8 (2-8-5)



Cisco와 Nvidia의 강력한 파트너쉽

- Cisco Secure AI Factory **with NVIDIA : March 18, 2025 발표 !**



Summary

- **고객의 비지니스 환경에 따라 선택할 수 있는 풀스택 AI 솔루션**

Small AI Cluster : **AI-POD**
Medium AI Cluster : **MGX 시스템 C845A**
Large AI Cluster : **HGX 시스템 C885A**
필요에 따라 네트워크, 컴퓨팅 및 스토리지를 독립적으로 확장할 수 있는 옵션

cisco + NVIDIA

- Cisco 는 NVIDIA의 Enterprise Reference Architecture 를 함께 합니다.
- Cisco 는 NVIDIA 와 함께 Cloud Partner Reference Architecture 를 만들어 갑니다.
- Cisco Silicon One® 과 NVIDIA Spectrum-X Silicon 을 통해 고객의 AI Journey 을 함께합니다.

Thank you