MakinaRocks

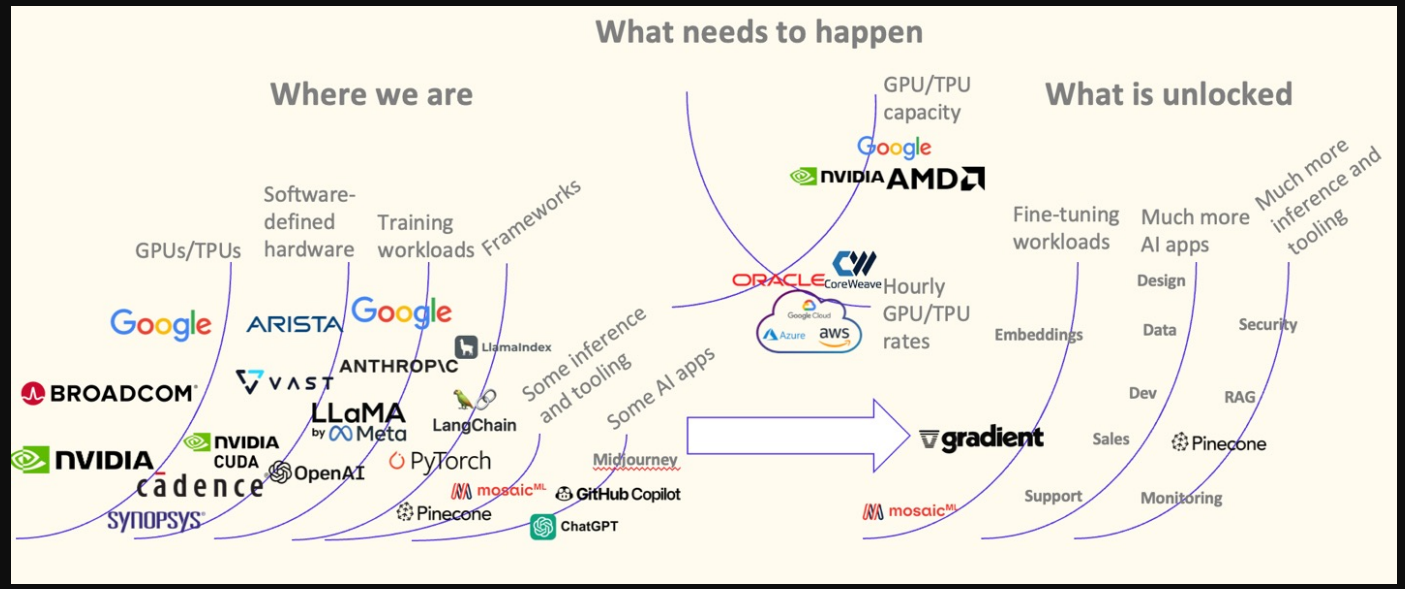
ATTENTION
AI Engineered for the Real World

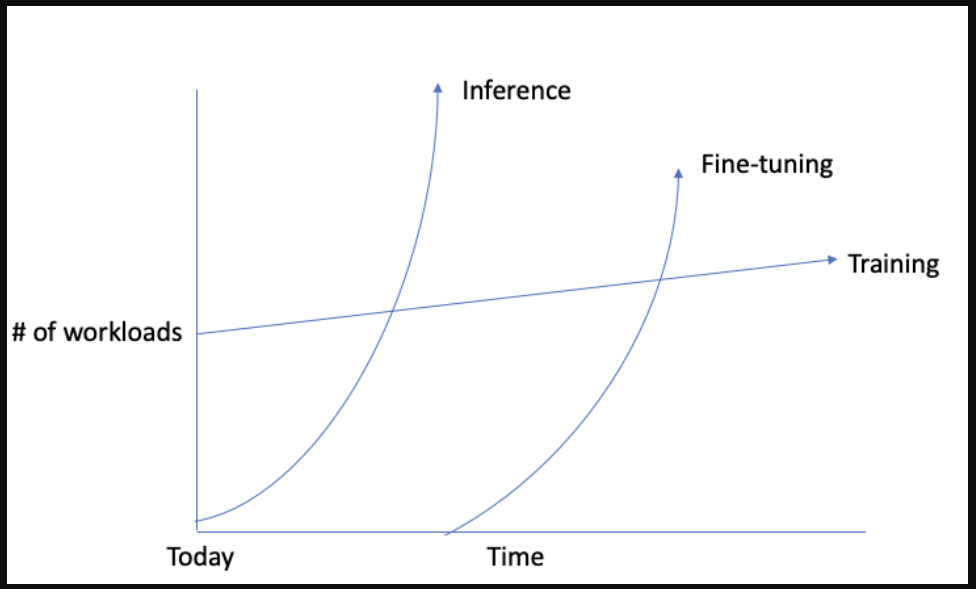
성공적인 엔터프라이즈 생성형 AI 여정을 위한 풀스택 인프라

2024

김진효 수석부장, 메이머스트

Industry Dynamics – AI 슈퍼 사이클이 온다





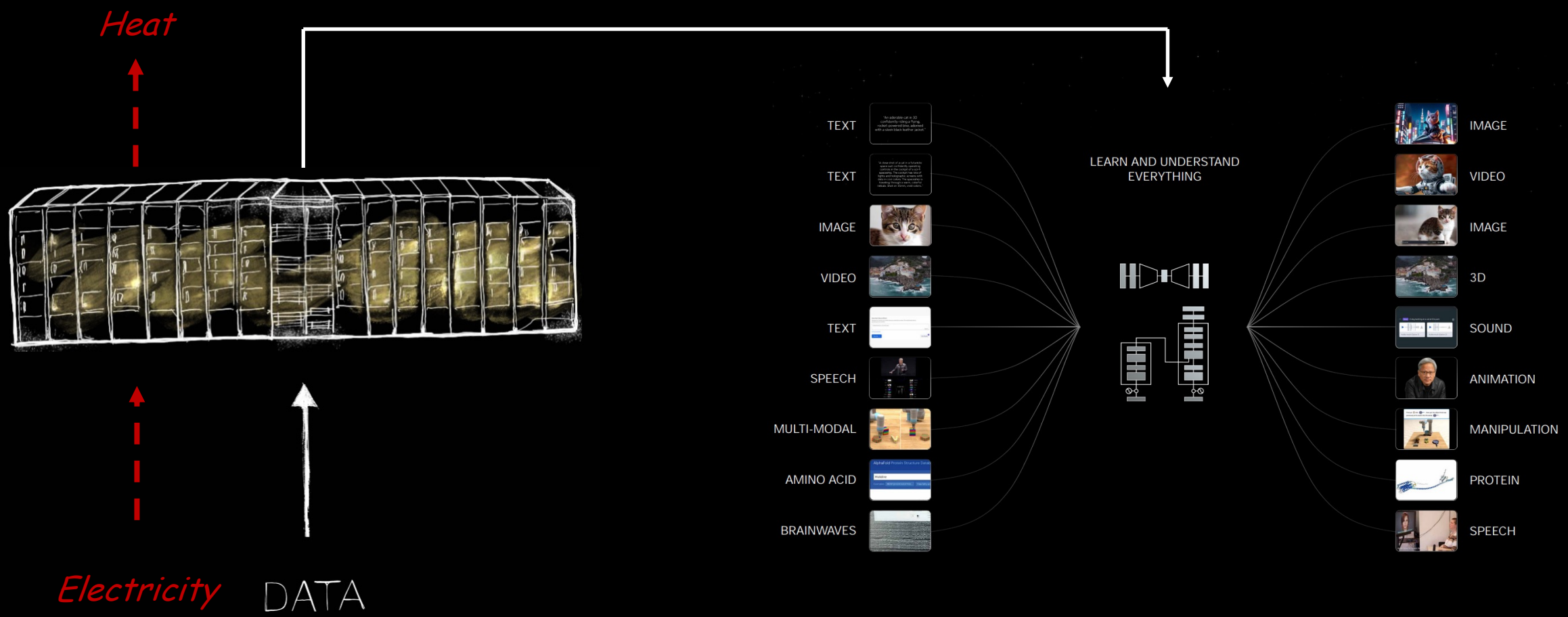
https://www.wing.vc/content/deep-dive-ai-workload-supercycle-nvidia-gpu-inference

AI 인프라 – Foundation Model, Data, Pre-Training

AI 서비스 – 선점 효과, AI응용/서비스, Fine-Tuning, Inferencing, 플랫폼

ATTENTION

AI 는 Data, Algorithms, 그리고 AI 인프라를 필요로 합니다

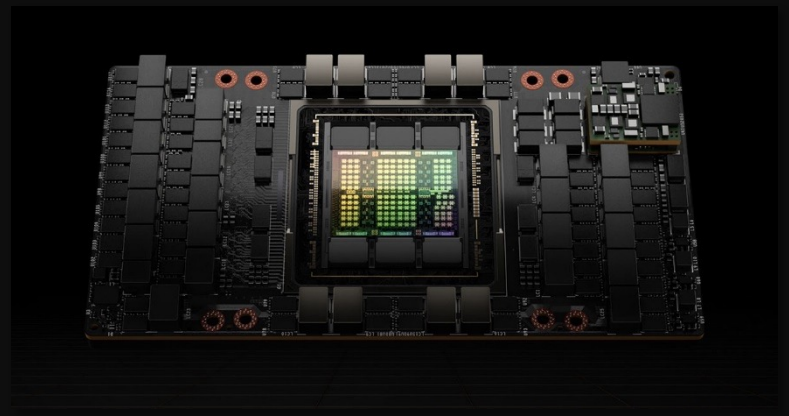


ATTENTION

Images source – NVIDIA presentations

Industry Dynamics – AI 인프라의 핵심은 GPU

AI 반도체, 네트워크, SW 생태계를 선점한 NVIDIA 지배력은 당분간 유지 전망



2023

2024

Bard

Feb, 2023: **Google** releases **Bard**

Meta

Feb, 2023: **Meta** releases its **LLaMA** family of **open-source models**

Claude

BY ANTHROPIC

March, 2023: **Anthropic**, a start-up founded in 2021 by ex-OpenAI researchers, releases **Claude**

GPT-4

March, 2023: **OpenAI** releases **GPT-4**

MISTRAL AI_ 7

Sept, 2023: **Mistral AI** releases **Mistral-7B** model

Grok

Nov, 2023: **xAI** releases **Grok**

Gemini

Dec, 2023: **Google** releases **Gemini**

Gemma 2

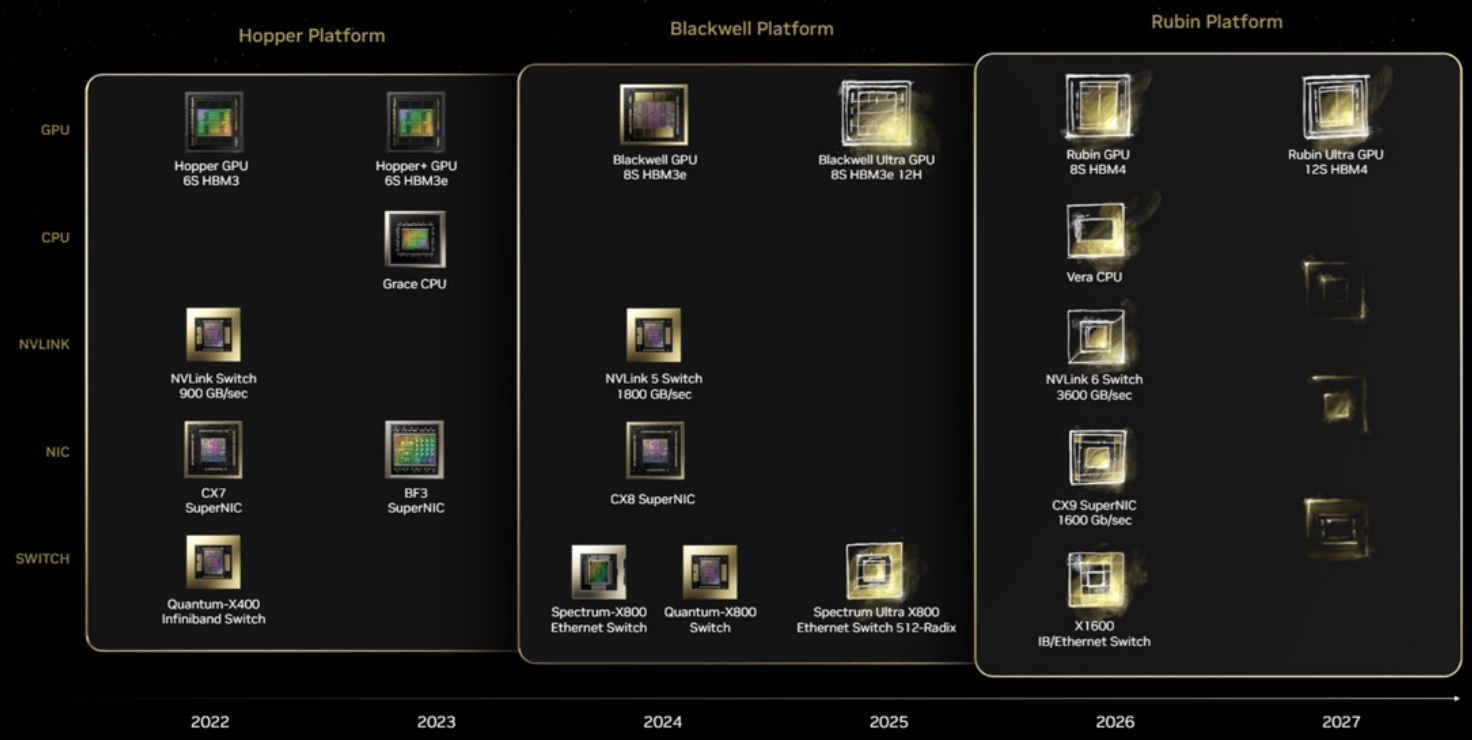
GPT-4o

Llama 3 From Meta

Codestral

Content Credit: https://lcs2-iitd.github.io/ELL881-AIL821-2401/static_files/presentations/101.pdf

AI Infrastructure is Evolving – (너무) 빠른 진화



DATACENTER SCALE • ONE-YEAR RHYTHM • TECHNOLOGY LIMITS • ONE ARCHITECTURE

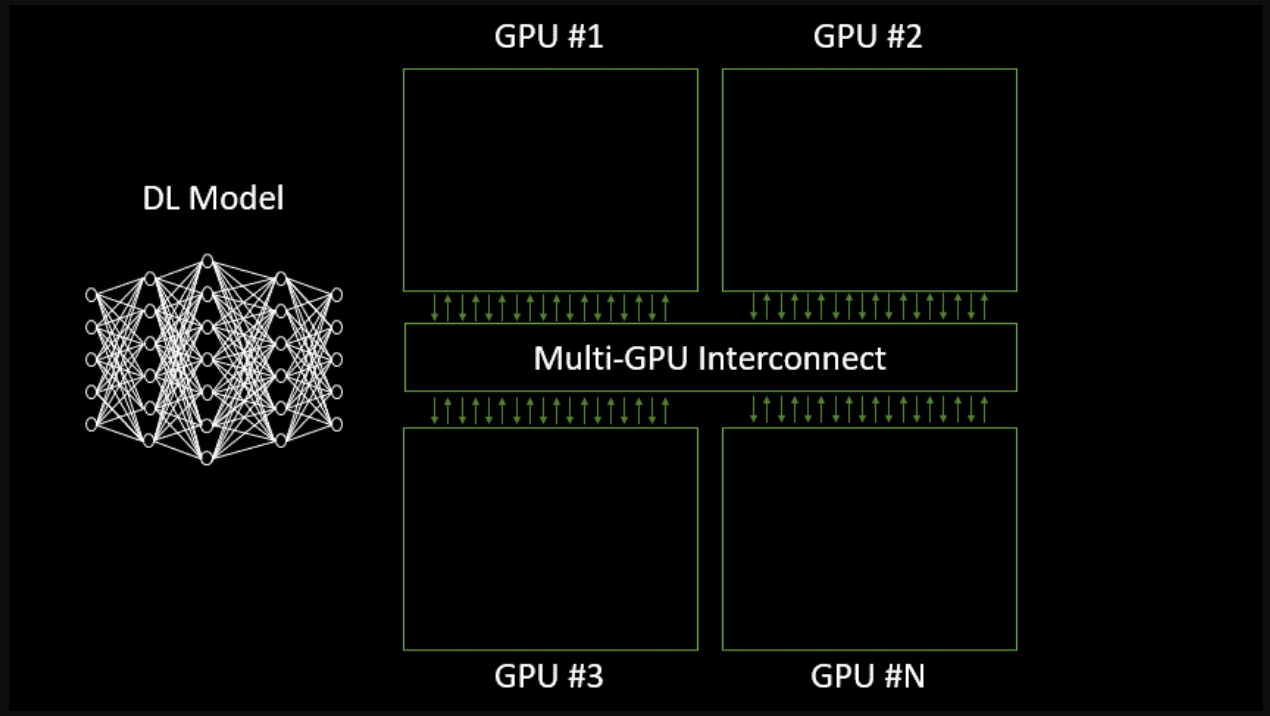
Images source – NVIDIA presentations

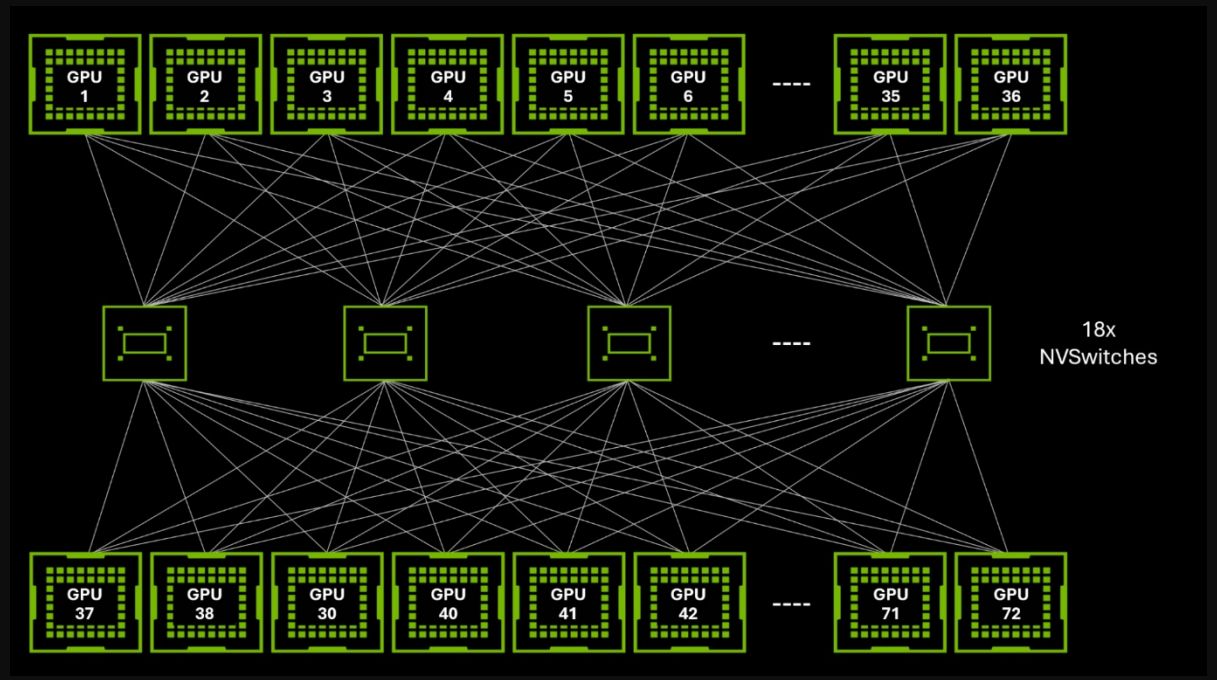
Rack-Scale Computing Platform

NVIDIA GB200 NVL72

ATTENTION

Multi-GPU Interconnect의 중요성





High volume of GPU-to-GPU communication happens with model parallelism

#N : 8 → #N : 72

High-Bandwidth Domain

Images source – NVIDIA Technical Blogs

전력과 열관리의 필요성

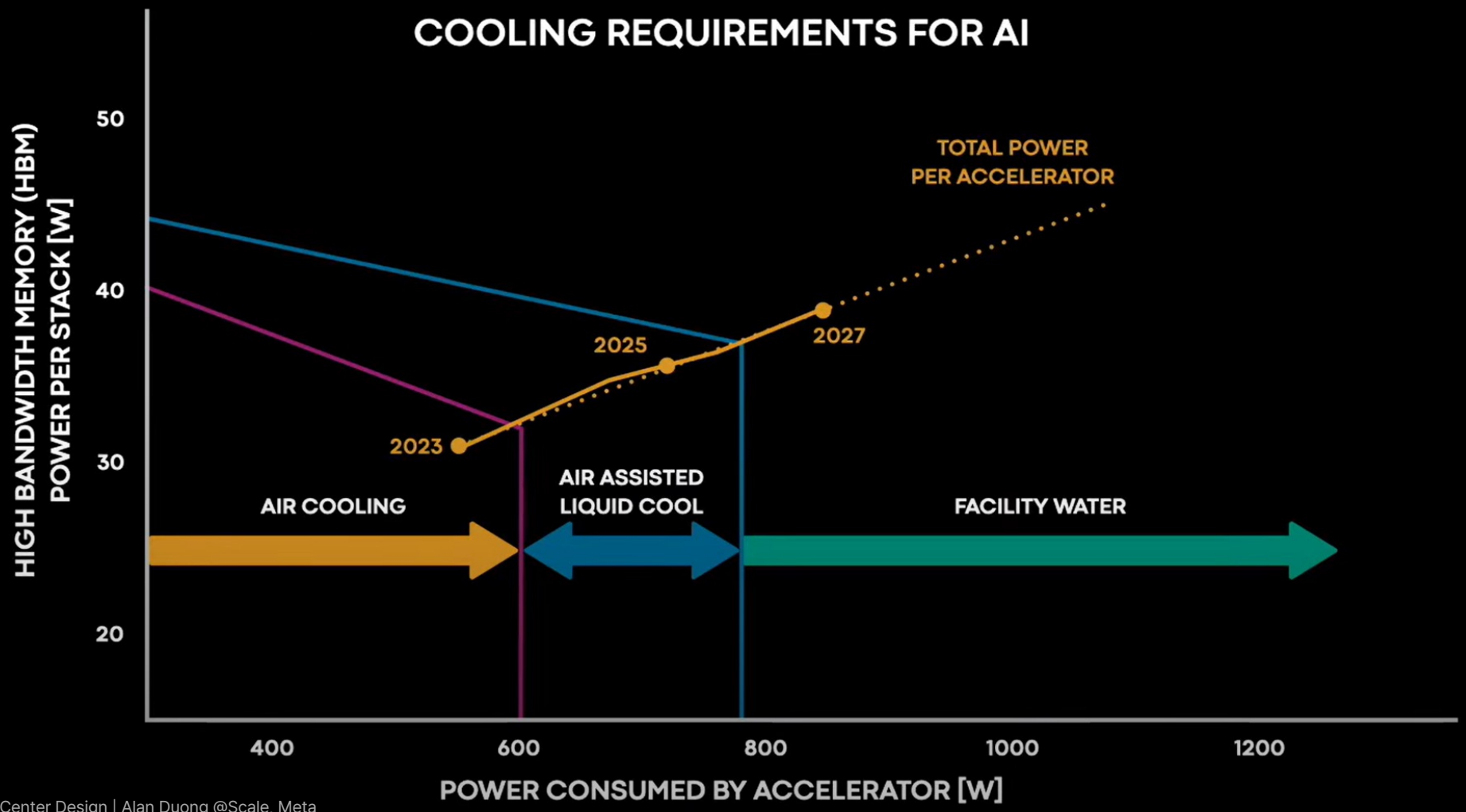
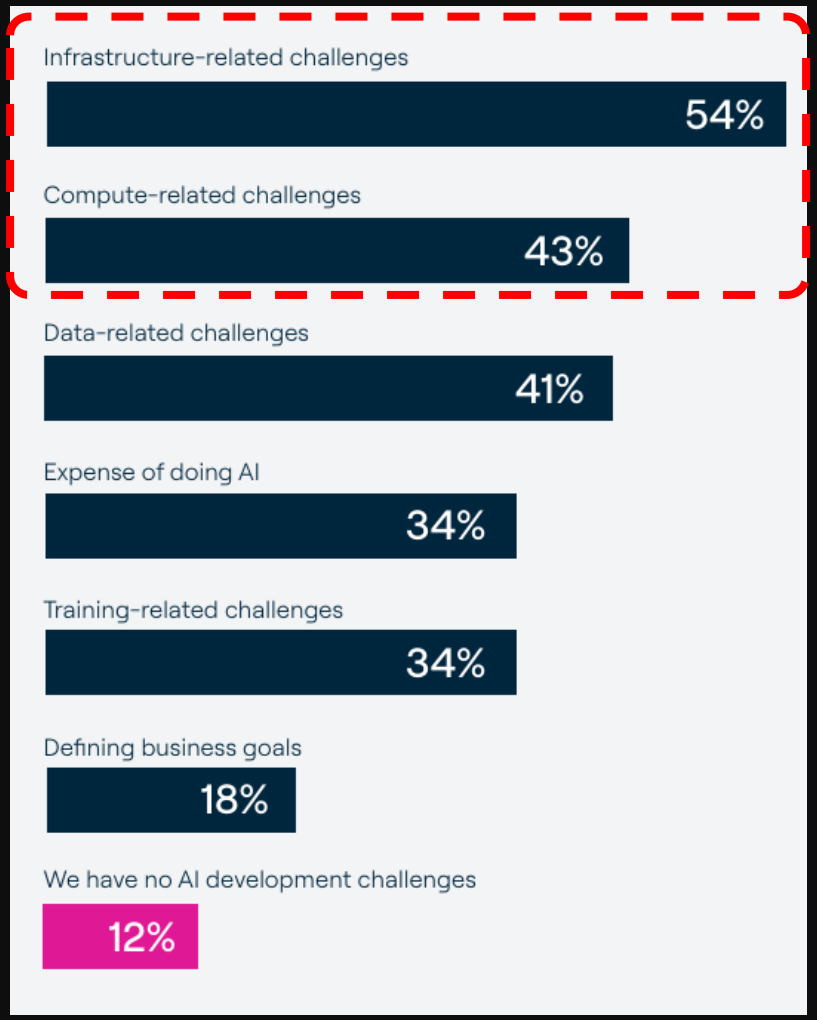


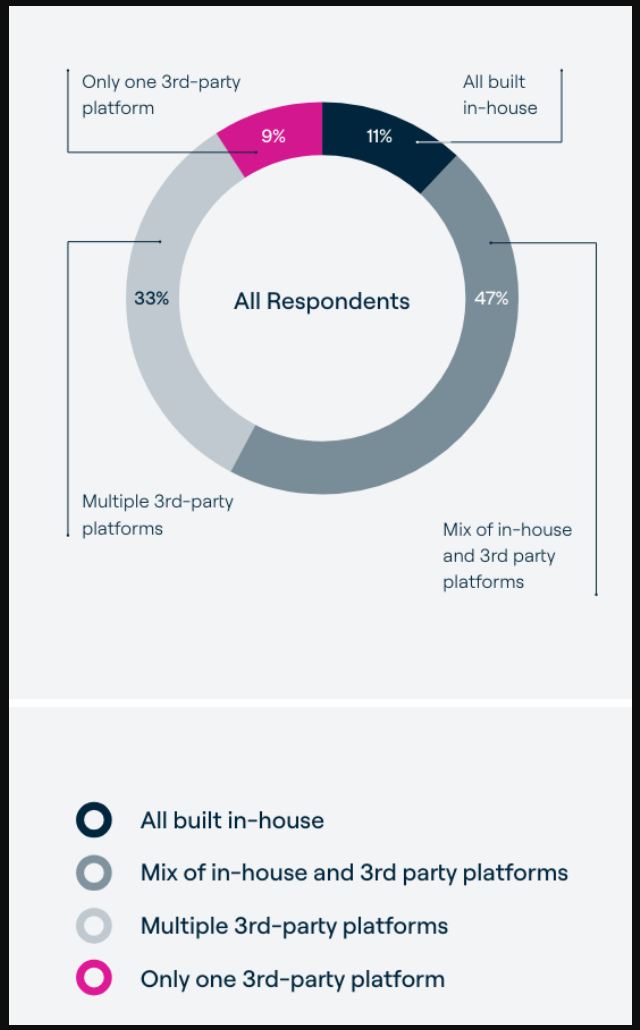
Image Source: Next-Generation Data Center Design | Alan Duong @Scale, Meta

AI 인프라에 대한 어려움/복잡성

Challenges for AI Development



AI/ML Stack Architecture



ATTENTION

The 2023 the State of AI Infrastructure Survey, run:ai

AI 인프라 (High Level) 네트워킹 모델

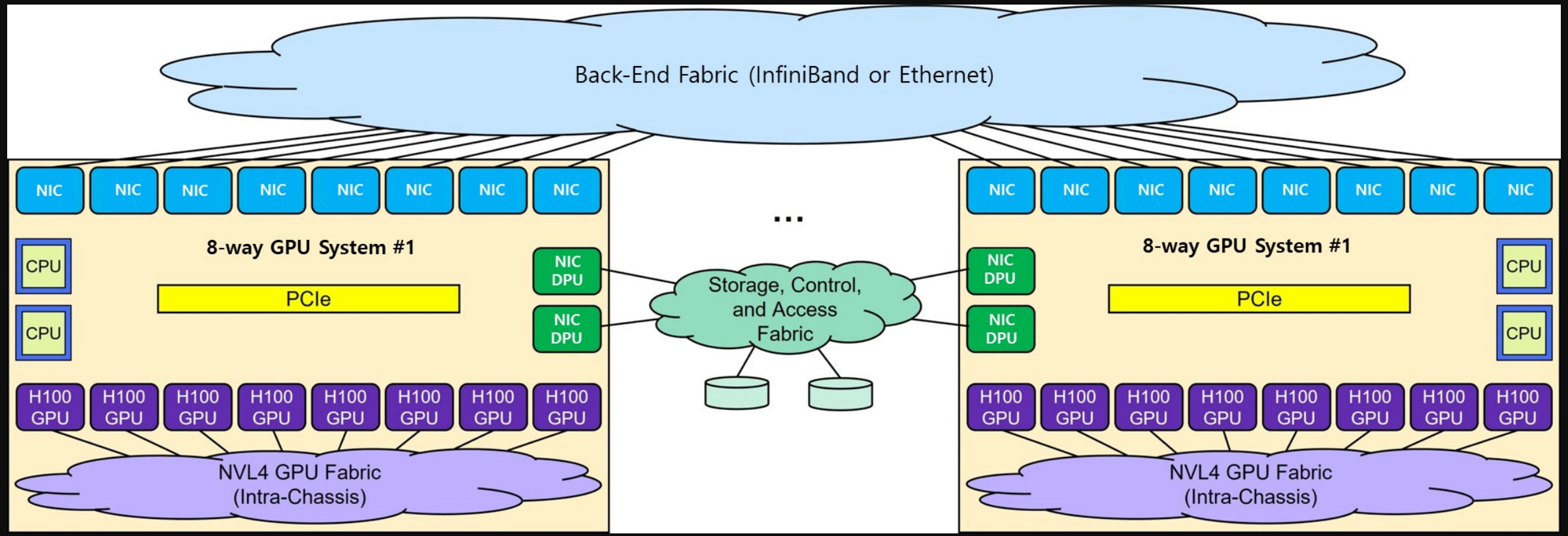
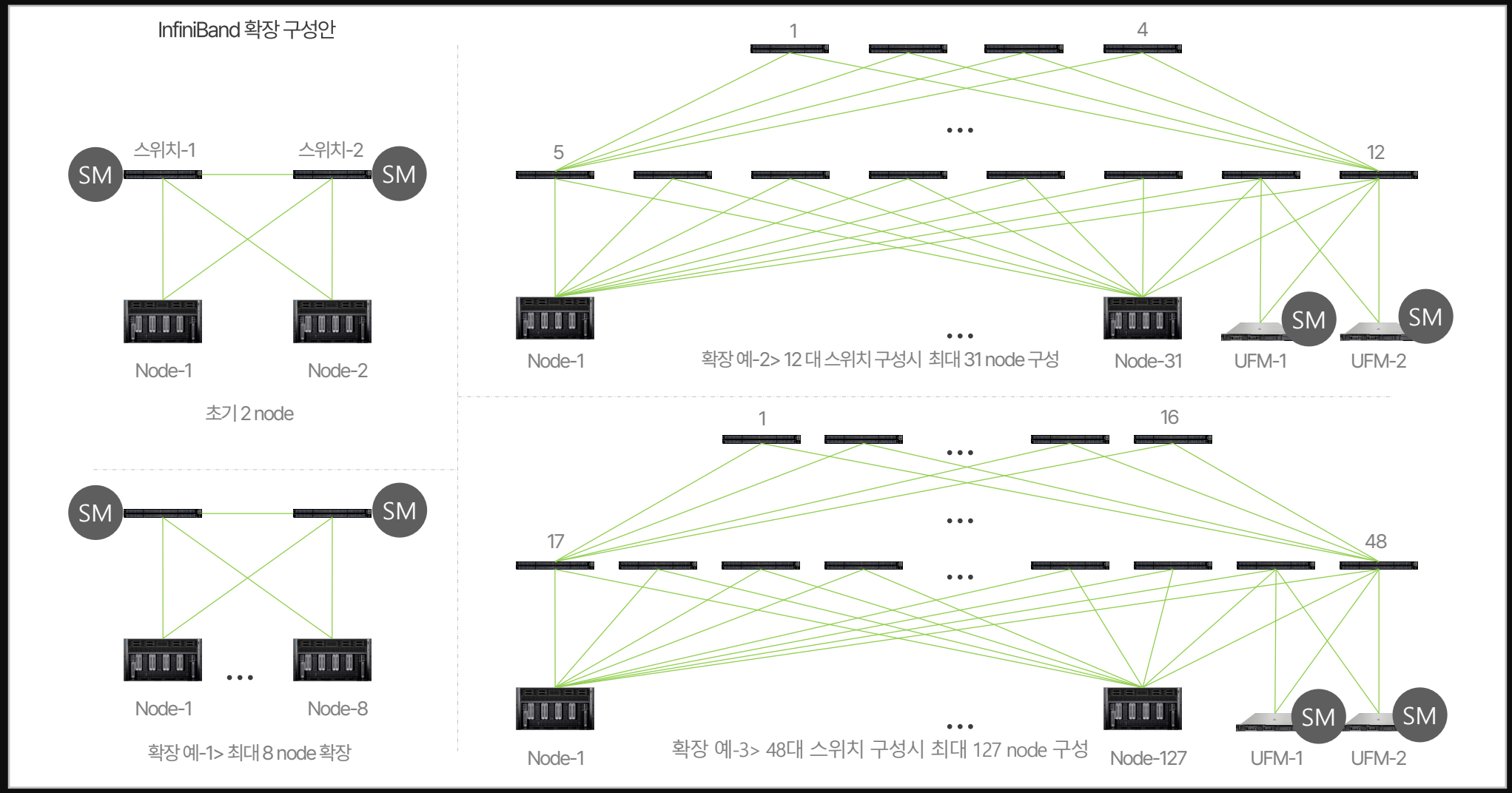


Image Source: Dell Technologies Tech. Blog: A-state-of-the-art-data-center-for-large-scale-ai

ATTENTION

멀티노드 GPU 클러스터 기본 디자인



- 서브넷 관리자(SM) – InfiniBand fabric 을 실행하고 관리하는 프로그램

ATTENTION

GPU 클러스터 검증

새로운 GPU 클러스터를 구축한 경우, 시스템에 스트레스를 가하여 모든 장비가 올바르게 작동하고 동일한 성능을 발휘하는지 검증하는 단계를 반드시 거쳐야 합니다.
DCGM 테스트, **NCCL** 테스트, **HPL/MLPerf** 벤치마크 테스트를 통해 클러스터의 각 노드가 동일한 수준의 계산/연산 성능 및 통신 성능을 수행하는지 검증합니다.

DCGM Test

- 개별 **Multi-GPU시스템 기능 및 성능**을 검증.

```
root@nccltest01:~#
root@nccltest01:~#
root@nccltest01:~#
root@nccltest01:~# dcgmi diag -r 4
Successfully ran diagnostic for group.
+---------------------------+------------------------------------------------+
| Diagnostic                | Result                                         |
+===========================+================================================+
|-----  Metadata ----------+------------------------------------------------|
| DCGM Version              | 3.3.5                                          |
| Driver Version Detected   | 545.23.08                                      |
| GPU Device IDs Detected   | 25b6,25b6                                      |
|-----  Deployment --------+------------------------------------------------|
| Denylist                  | Pass                                           |
| NVML Library              | Pass                                           |
| CUDA Main Library         | Pass                                           |
| Permissions and OS Blocks | Pass                                           |
| Persistence Mode          | Pass                                           |
| Environment Variables     | Pass                                           |
| Page Retirement/Row Remap | Pass                                           |
| Graphics Processes        | Pass                                           |
| Inforom                   | Pass                                           |
+---------------------------+------------------------------------------------+
root@nccltest01:~#
```

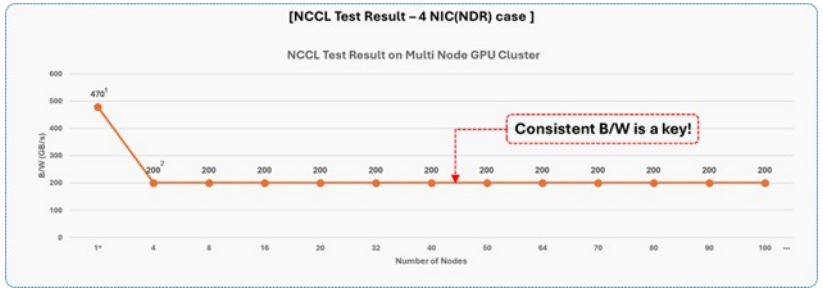
dcgmi diag –r 4 입력하여 하드웨어 진단 실시

```
root@nccltest01:~#
root@nccltest01:~#
root@nccltest01:~# dcgmi nvlink --errors -g 0
+-----------------------------+-----------------+
| NVLINK Error Counts         |                 |
| GPU 0                       |                 |
+=============================+=================+
+-----------------------------+-----------------+
root@nccltest01:~# dcgmi nvlink --errors -g 1
+-----------------------------+-----------------+
| NVLINK Error Counts         |                 |
| GPU 1                       |                 |
+=============================+=================+
+-----------------------------+-----------------+
root@nccltest01:~#
```

dcgmi nvlink –errors –g <gpu id> 입력하여 각 GPU의 nvlink 진단 실시

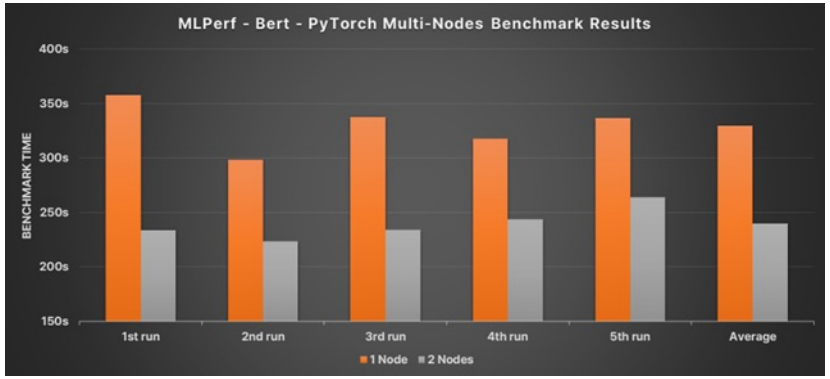
NCCL Test

- 노드 간 네트워크 대역폭이 예상대로 확보되는지 검증
- 노드 수가 증가함에 따라 네트워크 **대역폭이 일정하게 유지**되는 검증



HPL/MLPerf Test

- Multi-Node 환경에서 GPU Cluster의 계산/연산 성능이 예상대로 확보되는지 검증
- 노드 수가 증가함에 따라 **계산/연산 성능이 일정하게 증가**하는지 검증



ATTENTION

Dell AI Factory 플랫폼

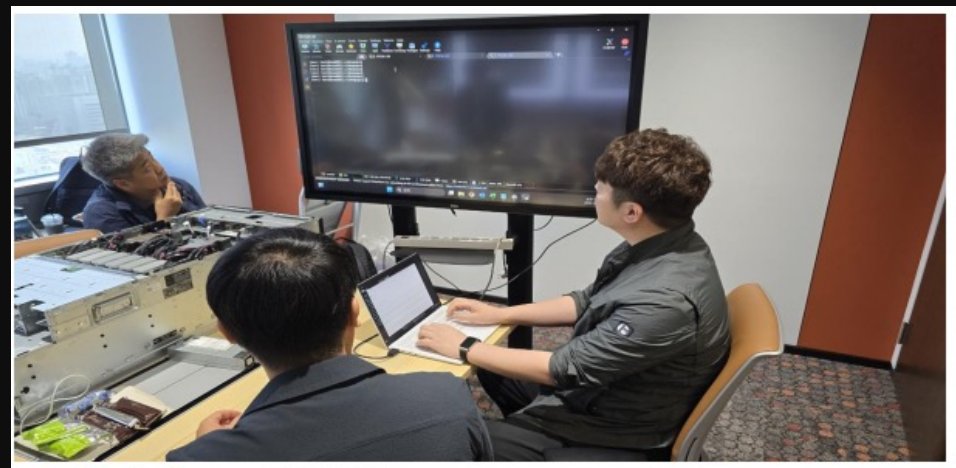
메이머스트는 고객의 성공적인 **AI 여정**을 위해 **Dell AI Factory** 플랫폼을 운영하고 있습니다.

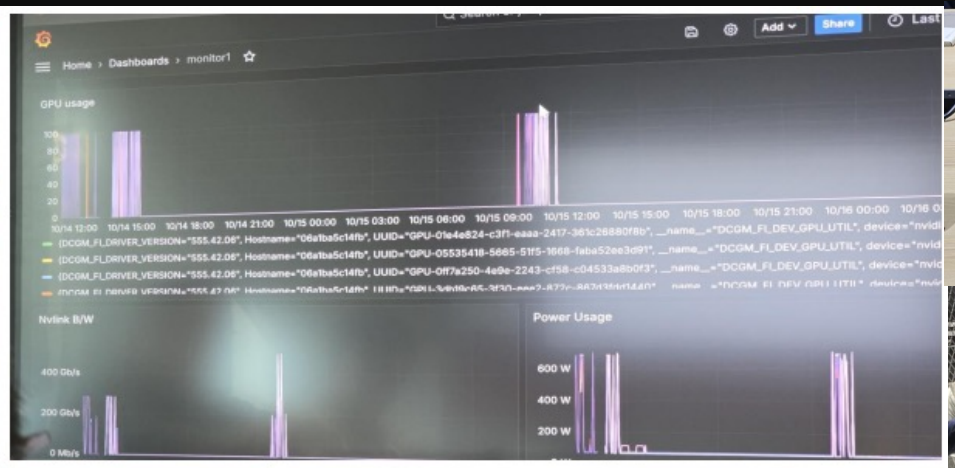
Dell AI Factory Demo/PoC Cetner

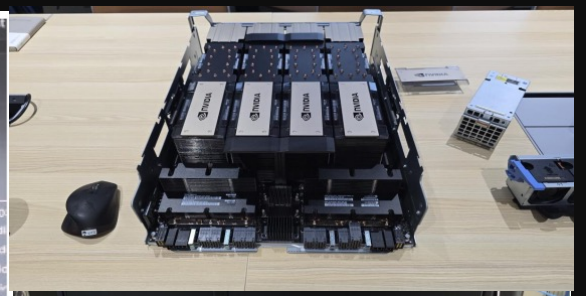
01	02	03	04
PoC/Demo Support	**Pre-Validations &Issues Handling**	**Optimization Test**	**Ecosystem Solutions Integration Test**
▪ Support Use-Cases Prototyping using Customer's Data Example) RAG Chatbot ▪ Workloads(Training/Inferencing) Test and Benchmark Test for Sizing	▪ Provisioning Test and Verification ▪ Slurm/k8s Cluster Implementation and Test ▪ Clustering Solution (BCM Essential) Test ▪ SW/Version Upgrade Test ▪ Issue Reproducing and Fixing	▪ NCCL-TESTS Test ▪ HPL Test Test ▪ MLPerf Test	▪ MLOps/LLMOps Solutions Pre-Integration and Test ▪ Domain Use-Cases Solution Pre-Integration and Test

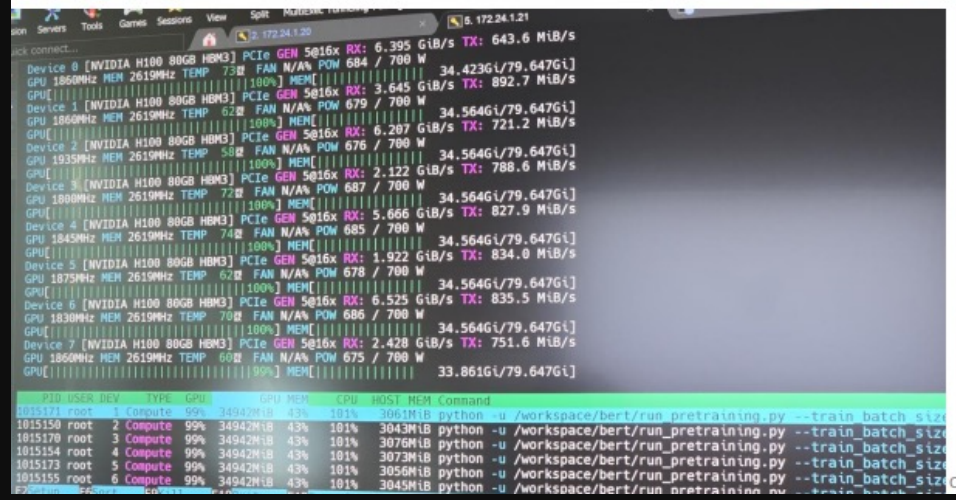
ATTENTION

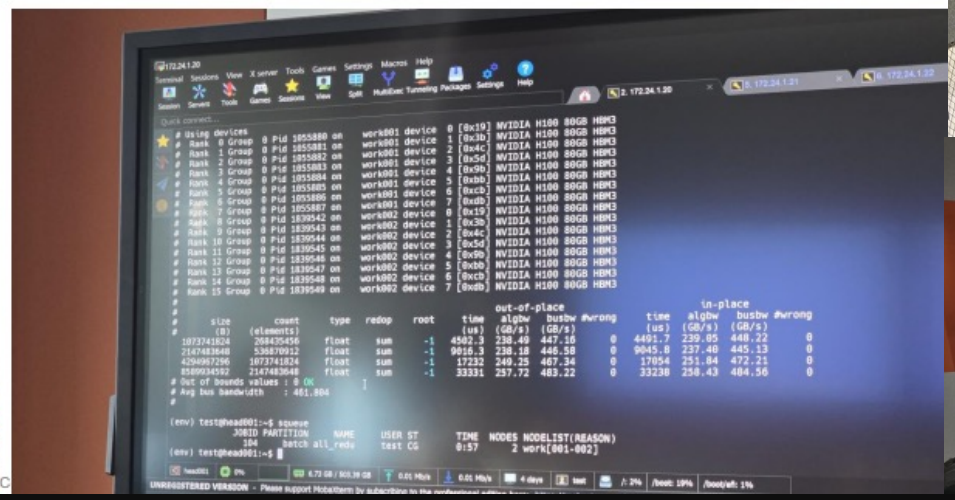
AI Factory 플랫폼을 통한 고객 기술 지원, 시연 및 Benchmark









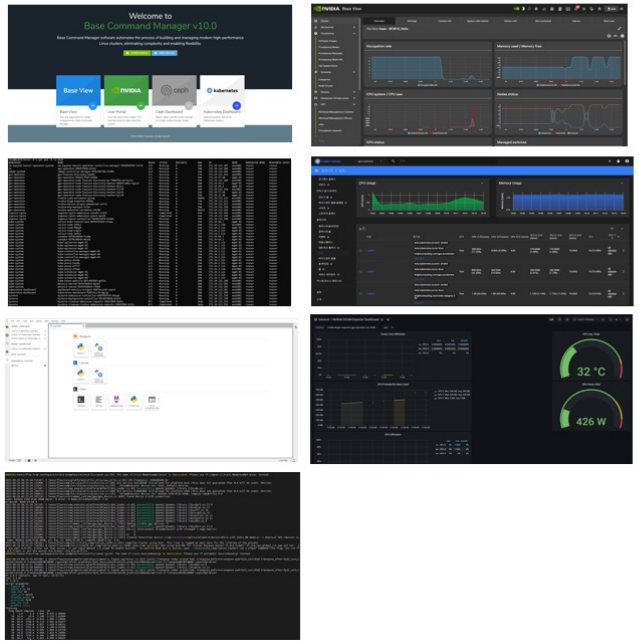


ATTENTION

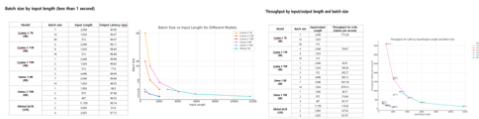
AI Factory 플랫폼 활용 예시– 1/3

NVIDIA BCM10

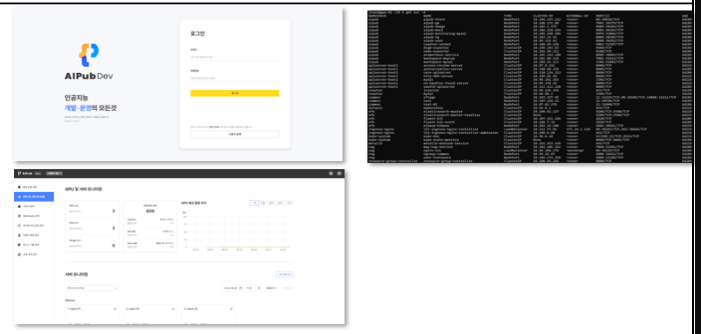
Test Plan	1. Environment configuration - Head Node (Dell PowerEdge 760) BCM installation and configuration - Provisiong Image Creation (Slurm, K8S) - Master Node (Dell PowerEdge 760 3nodes) configuration and Worker Node (Dell PowerEdge XE9680 2nodes) Image Provisioning 2. Slurm configuration and testing - Slurm configuration through CLI commands - Configuration verification and error checking - GPU Stress Test 3. K8S configuration and testing - Configure K8S with CLI commands - GPU Operator, Neteork Operator configuration - Jupyter HUB, Grafana, Prometheus, K8S Dashboard, Ingress Control configuration - Configuration verification and error checking - GPU Stress Test
Result	Nvidia bmc installation to ensure consistency of Dell infrastructure environments and consistency of key features related to k8 and GPUs 1. Slurm - No special configuration or errors - Verify normal operation of Multi GPU Stress 2. K8S - No special configuration or errors - Verify normal operation of Multi GPU Stress [Overall review] By configuring nvidia bcm, the customer can easily configure the necessary s/w stack, cluster the customer's gpu nodes, and monitor each node.



TensorRT-LLM Inferencing

Test Plan	1. Environmental Configuration - OS: Ubuntu 22.04 - BCM: 10.23.12 - Docker: 24.09 - Nvidia Container Toolkit: 1.16.0 - Cuda Toolkit: 12.2 - Model: Llama2 7B, Llama2 13B, Llama2 70B, Llama3 8B, Llama3 70B, Mistral 8*7B 2. Planning - Run TensorRT-LLM as Dockers - Check GPU memory for each - The same executable code applies different variables (8 GPUs, Batch Size, TP, InputLength, OutputLength). - TTFT, Latency, Throught Measurement 3. Reference URL - https://github.com/NVIDIA/TensorRT-LLM/blob/v0.8.0/docs/source/performance.md
Result	 The batch size roughly doubles as the input length decreases by ½. As the input/output length decreases by ½, the batch size roughly doubles and the throughput roughly increases by 1.6-2.8x

ISV Solutions

Test Plan	1. Environment configuration - OS: RHEL 8.7 - K8S: 1.29 - Master Node (Dell PowerEdge 760 3nodes), Work Node (Dell PowerEdge XE9680 2nodes) 2.Plan . 고객요구 사항 정의 . 사용자 계정, LB VIP, DNS, 스토리지 및 컨테이너 레지스트리 - Ten AI PUB K8S Deploy - Ten AI PUB Core Engine Deplop
Result	 Check out key features of K8, including consistency and GPU block splitting capabilities for Dell infrastructure environments. [Overall review] ai pub dev is composed of a UI so administrators can conveniently allocate resources. Multiple teams share resources and provide workload and monitoring capabilities. Its strength is that it provides advanced GPU management (GPU block splitting) functions and can split up to 1 GPU into 100.

AI Factory 플랫폼 활용 예시– 2/3

Training /w NVIDIA Nemo FW

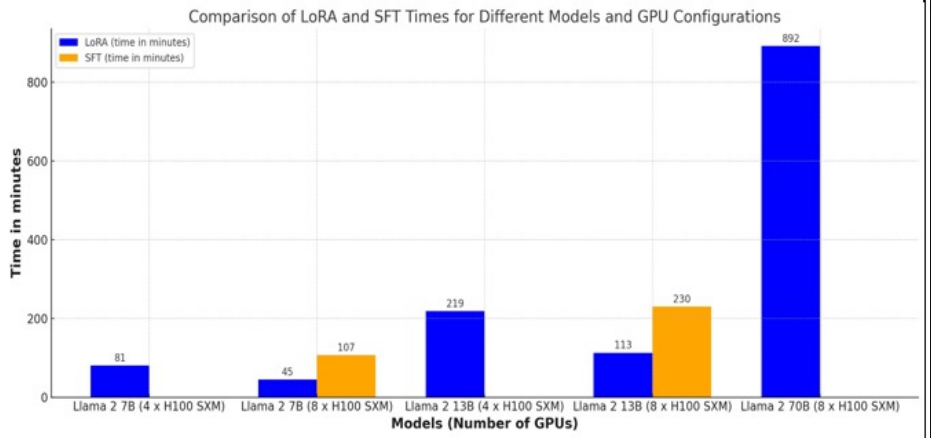
Test Plan

1. Environment configuration
- OS: Ubuntu 22.04
- BCM: 10.23.12
- K8S: 1.27.13
- NeMo Framework Training container: 24.07
- GPU Operator: 24.6.2
- Network Operator: 23.5.0
- DataSet: databricks-dolly-15k
- Llama 2 Model Download and convert model to NeMo

2.Plan
- Measurement of Fine-Tuning (SFT, Lora) time for Llama 2 models (7B, 13B, 70B) according to GPU quantity (4 GPU, 8 GPU)
- Llama 2 7B, 4 GPU, Lora
- Llama 2 7B, 8 GPU, (Lora, SFT)
- Llama 2 13B, 4 GPU, Lora
- Llama 2 13B, 8 GPU, (Lora, SFT)
- Llama 2 70B, 8 GPU, Lora

3. Reference URL
- https://www.delltechnologies.com/asset/en-us/solutions/infrastructure-solutions/technical-support/gen-ai-model-customization-design-guide.pdf

Result



[Overall review]
Fine tuning (Lora, SFT) time was measured for Llama 2 7B, 13B, and 70B models with XE9680 H100 8 GPU.
Fine tuning from 7B to 70B (excluding SFT) is possible with XE9680 1 Node.

RAG prototyping /w Customer's data

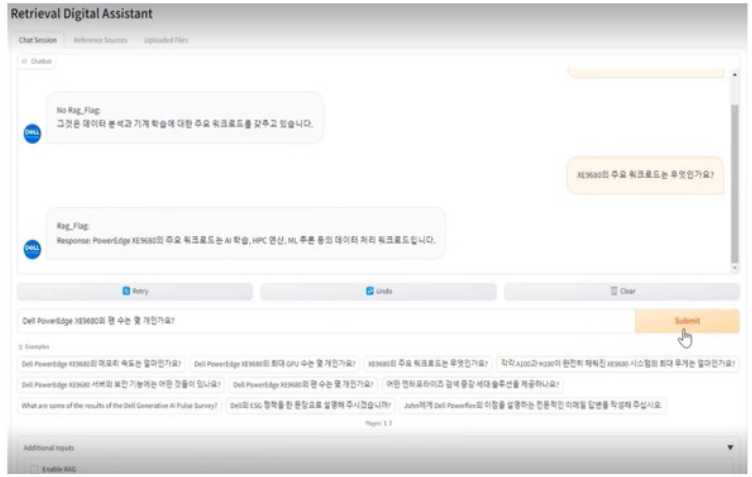
Test Plan

1. Environment configuration
- OS: Ubuntu 22.04
- BCM: 10.23.12
- K8S: 1.27.13
- GPU Operator: 24.6.2
- Network Operator: 23.5.0
- DataSet: XE9860, XE8640 technical-guide
- Pre-trained Model: Mistral 7B instruct

2.Plan
- K8S Rag Chatbot Container configuration
- Data set creation and labeling in XE9860, XE8640 technical-guide
- Fine tuning (check accuracy) with Mistral 7B model data set
- Writing Huggingface Pipeline Code
- Code customization

3. Reference URL
- https://github.com/dell-examples/generative-ai

Result

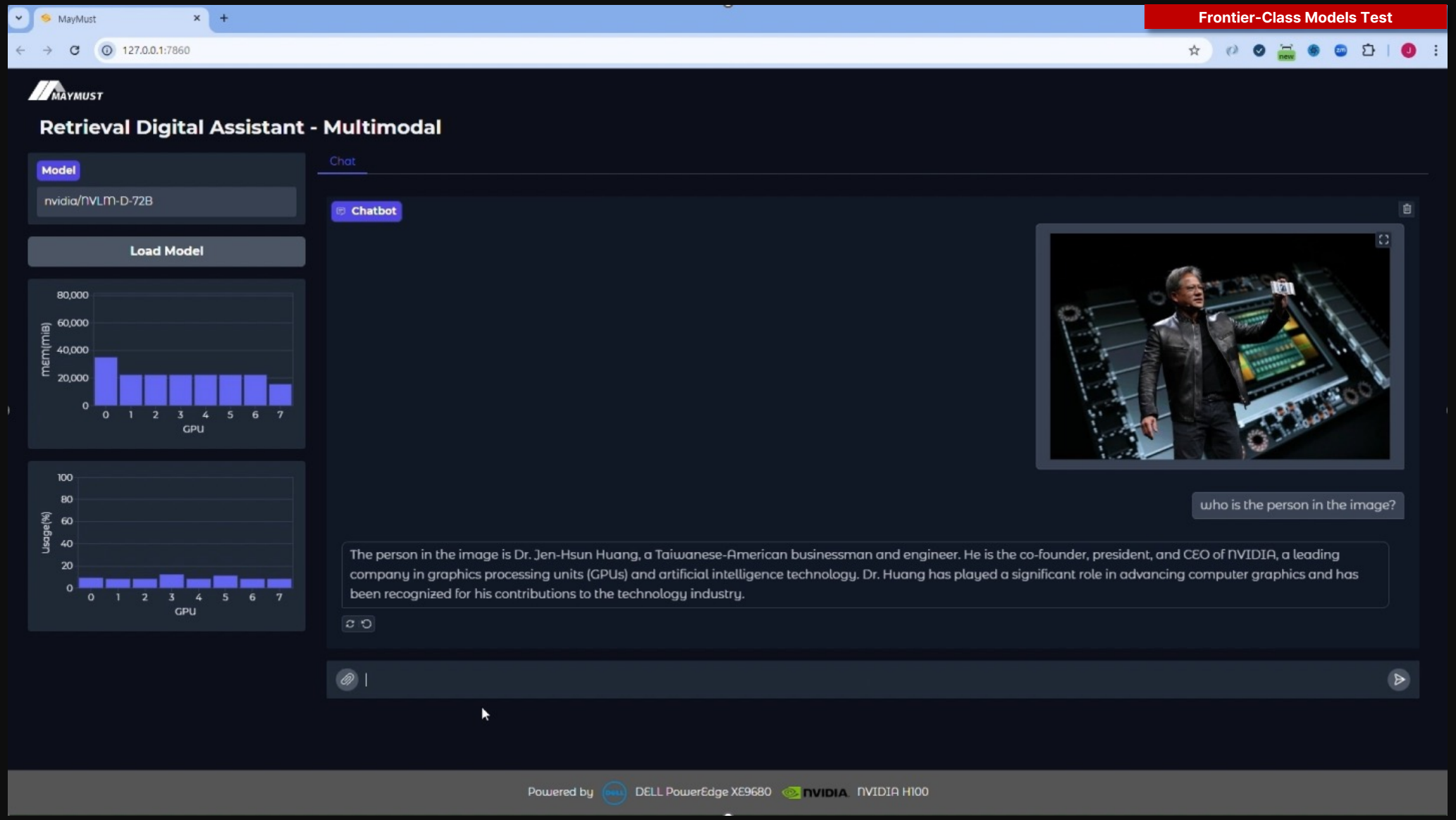


When asking the chatbot about servers on xe9680, 8 out of 10 questions were answered correctly.

[Overall review]
After selecting a base model with good performance, creating a dataset of the customer's domain area and the customer's in-house data, and then fine tuning the model, you can configure a RAG chatbot specialized for the customer.

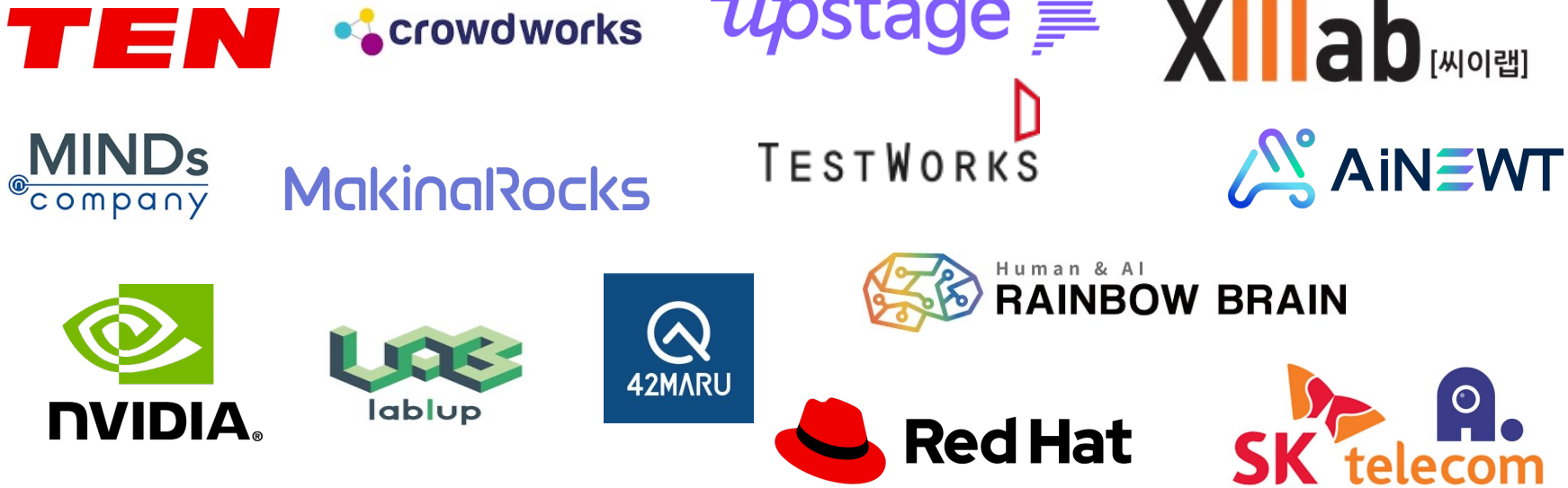
ATTENTION

AI Factory 플랫폼 활용 예시 – 3/3



AI Full Stack을 위한 에코시스템

AI Enterprise Ecosystem Solutions



Optimized & Validated GenAI Full-Stack Infra.

MAYMUST

고객의 빠른 성공을 위해, AI Full Stack을 제공해 드립니다.

메이머스트의 최우선 목표는 고객의 성공입니다.
업계 최고의 에코시스템 파트너사와 함께 고객의 다양한 AI 워크로드 및 Use-Case를 최적의 인프라 환경하에서 지원합니다.

Full Stack Offering from Infra, SW to Use-case

Use-Cases

Integrated Proposal Service with AI Ecosystem partners

- SKT Enterprise
- MNC, MakinaRocks, AiNEWT

MLOps/LLMOps

Pre-installed and tested MLOps/LLMOps deployment service

- (Lablup)
- TEN
- SKT Enterprise

Optimized AI SW

AI SW technical support service

- NVIDIA optimized AI SW installation & tech support
- GPU Cluster deployment
- NVIDIA and Dell Technologies SW Solutions deployment & tech support
- Containerized workloads support

GPU Optimized Infra

Expert Groups for AI Infrastructure design & implementation

- Optimized AI fabric design & implementation service
- End-to-End Systems (Compute, Network, Storage) design, implementation, and validation (DCGM test, NCCL test, and HPL/MLPerf benchmark test) Service

THANK **YOU!**