HS 효성인포메이션시스템 | Hitachi Vantara

[2025 데이터센터 서밋 코리아]

AI 데이터센터로의 발전:

에너지 효율과 성능을 모두 잡는 최신 인프라 전략

강준범 컨설턴트
HS효성인포메이션시스템
2025년 7월 14일

Agenda

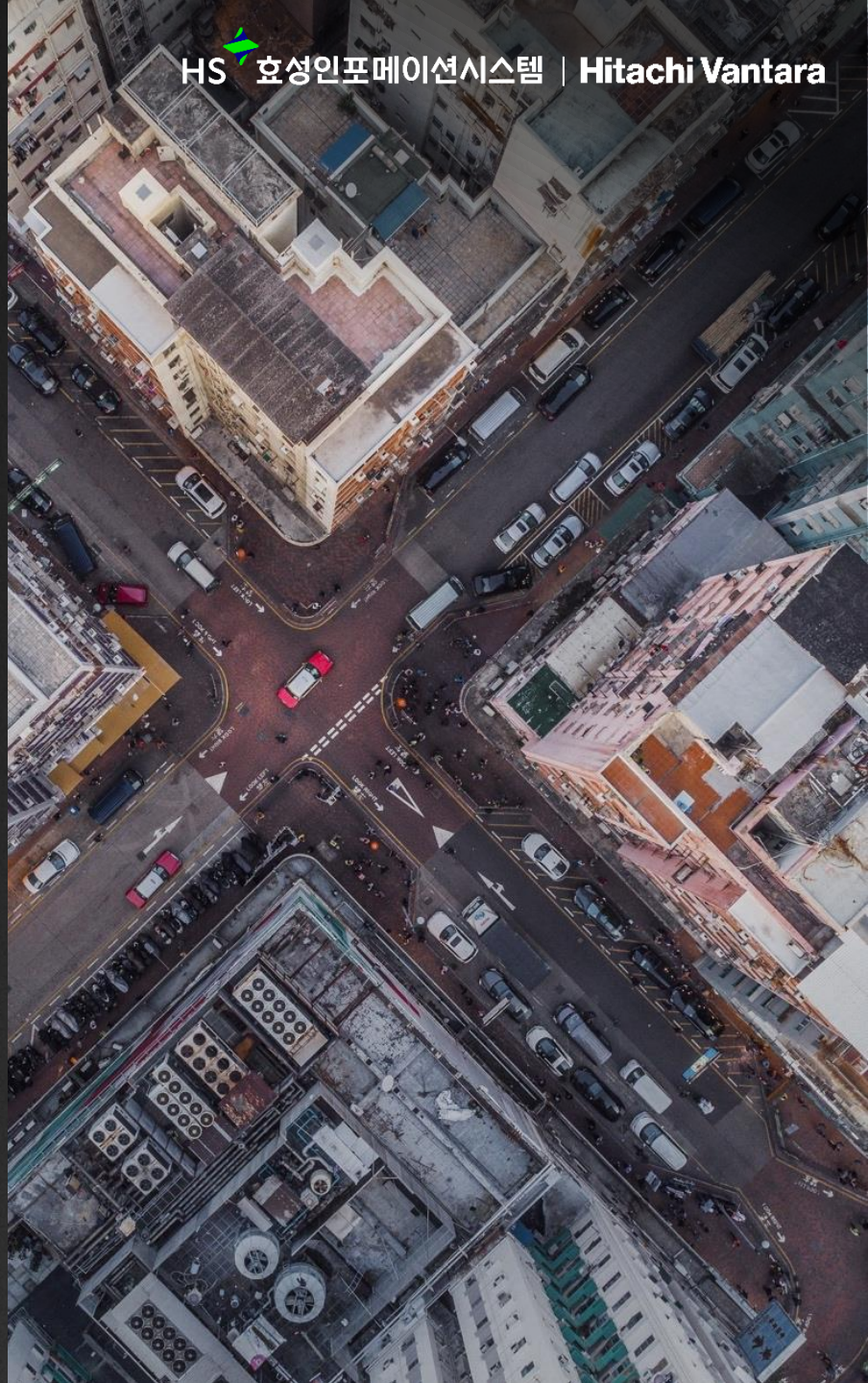
I. 데이터센터 트렌드

II. HS효성 AI 플랫폼

III. 대표 구축 사례

I. 데이터센터 트렌드

1. 데이터센터 등장
2. AI 데이터센터의 필요성
3. AI 시대, ’25 AI Keyword
4. 기존 데이터센터 vs. AI 데이터센터
5. AI 데이터센터의 주요 구성 요소
6. AI 데이터센터 구성의 복잡성



데이터센터 등장

- 고성능을 요구하는 AI 인프라를 수용하기 위해 AI 데이터센터 등장

자체 컴퓨터 배치

- ✓ 자원의 효율적 관리?
- ✓ 비효율적인 공간 차지?
- ✓ 중복 비용 투자?

데이터센터 등장

- ✓ 컴퓨팅 자원의 효율적 관리
- ✓ 안정적인 서비스 제공
- ✓ 비용 절감
- ✓ IT 인프라의 집중화

AI 데이터 센터 등장

- ✓ AI 기술의 급격한 발전과 대규모 AI 모델 학습·추론 수요 증가에 대응하기 위해 등장

AI 데이터센터의 필요성

- AI 데이터센터는 정부와 민간기업 모두가 **미래 산업의 핵심 인프라**로 주목하는 국가 전략 프로젝트
- 생성형 AI의 급성장, 기존 데이터센터로는 감당할 수 없는 **막대한 연산**과 **데이터 처리**를 위해 **AI 데이터센터** 등장

경제

[단독] AI데이터센터도 반도체처럼 稅혜택... 국가전략시설 지정해 최대 25% 공제

김정환 기자 flame@mk.co.kr
류영욱 기자 ryu.youngwook@mk.co.kr
입력 : 2025-06-18 18:11:47 수정 : 2025-06-18 19:00:01

국정기획委, 부처별 업무보고

100조원 규모 전략투자 구체화
50조 규모서 연기금 등 증액도

최신뉴스

민간에 힘실은 AI정책…국가데이터센터 'N개' 지원 부상

송고 2025-06-22 07:00

조성미 기자
+ 구독

AI 100조 투자 민관 협력 불가피…'AI 고속도로' 해외 빅테크도 참여할 듯
현행 과기정통부 체제론 AI 총괄 거버넌스 어려울 듯…'부총리 격상·AI 집중' 거론

(서울=연합뉴스) 조성미 기자 = 이재명 정부가 국가 인공지능(AI) 정책을 민간
한 투자에 방점을 찍으면서 윤석열 정부에서 관 주도 성격이 짙었던 AI 인프라
직한 변화가 예상된다.

새 정부가 민간 주축의 AI 정책으로 선회한 배경에는 막대한 투자 시급성에다
정책을 주도한 국가인공지능위원회가 업계보다 학계 의견 위주로 목소리를 내
어졌다는 지적이 반영된 것으로 보인다.

정치 > 정치일반

"AI데이터센터, 차세대 국가 SOC 사업"

서지윤 기자
파이낸셜뉴스 입력 2025.06.23 18:19 수정 2025.06.24 08:07

국정위, 세액공제 등 지원 구체화
공공데이터 개방해 AI 활용 높여

국정기획위원회가 이재명 정부의 '인공지능(AI) 3대 강국' 공약의 밑그림을 그릴 TF를 출범한 가운데 AI 산업 육성의 핵심인 데이터 인프라 및 공급 방안에 이목이 쏠린다. 국정위는 민간 중심의 데이터센터 생태계 구축과 공공데이터 개방에 중점을 두고 정책 초안을 마련할 것으로 보인다.

23일 국정기획위 '대한민국 진짜성장을 위한 전략 보고서'에 따르면 AI 3대 강국으로 도약하기 위한 주요 방안으로 AI 데이터센터 확대가 거론된다.

AI 학습을 위해선 방대한 데이터를 저장·처리·분석할 수 있는 고성능 컴퓨팅 자원과 안정적인 저장공간인 데이터센터가 필요하다. 이 정부는 AI 데이터센터를 차세대 국가 사회간접자본(SOC)으로 규정하고 최신 그래픽처리장치(GPU)를 확보한 AI 데이터센터 건설을 지원하겠다는 방침이다.

사회 환경

하정우 AI수석, "데이터센터에 재생에너지 공급, 정부가 주도해야"

인공지능 혁신성장을 위한 에너지정책방향 국회 토론회
"용인반도체클러스터, 수도권 전기 병목 키울 것"

옥기원 기자
수정 2025-06-18 18:54 등록 2025-06-18 18:49

기사를 읽어드립니다 5:34

전 세계 데이터센터들이 알이(RE)100 달성을 위한 재생에너지 확보 경쟁에 나섰다. 사진은 풍력단지 옆에 건설된 구글 데이터 센터 전경. 구글 누리집 갈무리

"전력 소비가 작건 커지건, 모든 경우에 재생에너지는 2035년까지 데이터센터의 추가 전력 수요 대부분을 공급할 것이다."

AI 시대, ’25 AI Keyword

개인 성향, 관심,
업무에 맞는
AI 맞춤 초개인화 경험

초개인화 AI

개인화된
AI Agent를 통한
단순 작업 자동화에서
복잡한 업무까지 수행

AI Agent

데이터 주권과
기술 자립 확보를 위해
독립적으로
AI를 구축·운영

Sovereign AI

기존 데이터센터 vs. AI 데이터센터

기존 데이터센터		AI 데이터센터
일반 컴퓨팅 (웹, DB, 클라우드, 엔터프라이즈 앱 등)		AI/ML 특화 (딥러닝, 자연어 처리, 이미지 인식 등)
5~15 kW/랙		40~120 kW/랙 (최신 AI 기준)
Air Cooling (공냉식)		Liquid Cooling (액체냉각 – 직/간접, 칩, 침지 등)
PUE* 1.5 ~ 2.0 (일반적) 일부 구형/소규모는 2.5 이상		PUE 1.1 ~ 1.5 (최신/대형 데이터센터 기준) 1.0 근접을 목표로 지속적 개발 진행

** PUE(Power Usage Effectiveness, 전력 사용 효율성) = 데이터센터 전체 전력 소비량 ÷ IT 장비 전력 소비량*

AI 데이터센터 주요 구성 요소

AI 데이터센터

고성능 컴퓨팅 (HPC)

Nvidia GPU /
AMD GPU NPU 등

고성능 데이터 저장소

고성능 AI 분석을 위한
효율적인
고성능 스토리지

Resource 효율화 솔루션

GPU 자원 효율을
높이기 위한
AIOps Stack

ESG를 위한 전력 효율화 솔루션

Arm 기반 친환경 서버
Liquid Cooling 냉각 방식

AI 데이터센터 구성의 복잡성

AI 데이터센터 설계 시 HPC 클러스터부터 고성능 스토리지·GPU 활용도까지, **복합적인 HW 및 솔루션 구성에 대한 검증** 필요

→ Reference 기반 **최적의 구성안 설계** 필요!

이슈 1. AI 솔루션 기술 부족

- AI플랫폼은 복잡한 인프라 및 솔루션 조합으로 구성
  (모델링 알고리즘, 클라우드, 컨테이너, GPU/서버가상화)

이슈 2. 초기 투자 비용 부족

- H/W 인프라에 더해 AI 솔루션에 대한 비용 부담, BigBang 형태의 투자에 대한 부담감
  (서버, 스토리지, 네트워크, AI/ML Ops 솔루션과 구축비용)

이슈 3. 전문 인력 및 역량 부족

- 기업내 내부 AI역량 부족에 대한 우려, 역량 있는 AI 파트너사 중요
  (구축 및 안정적 운영을 위한 기업내 AI역량 확보 이슈)

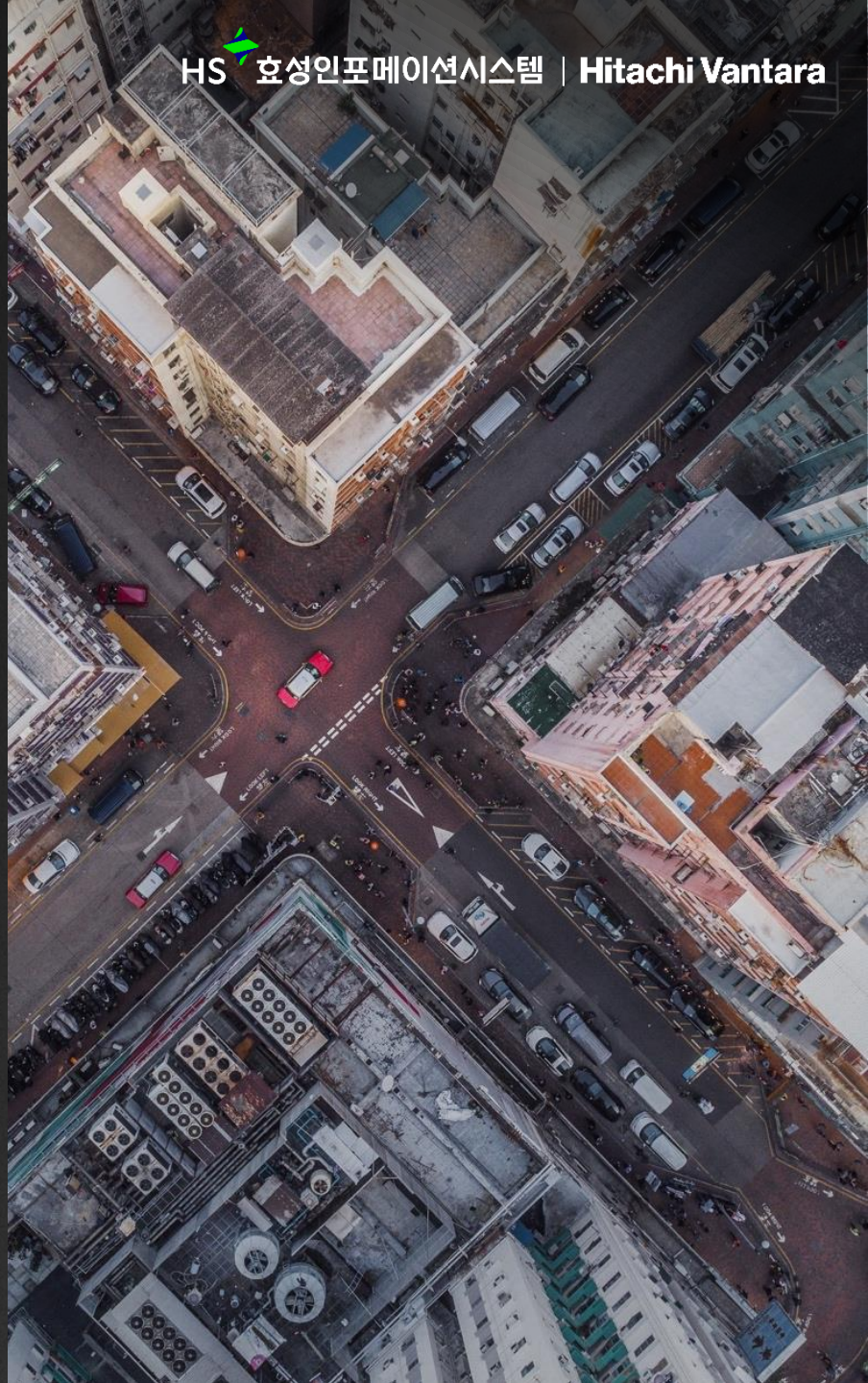
AI 시작은?

도입 후 활용은?

어떻게?

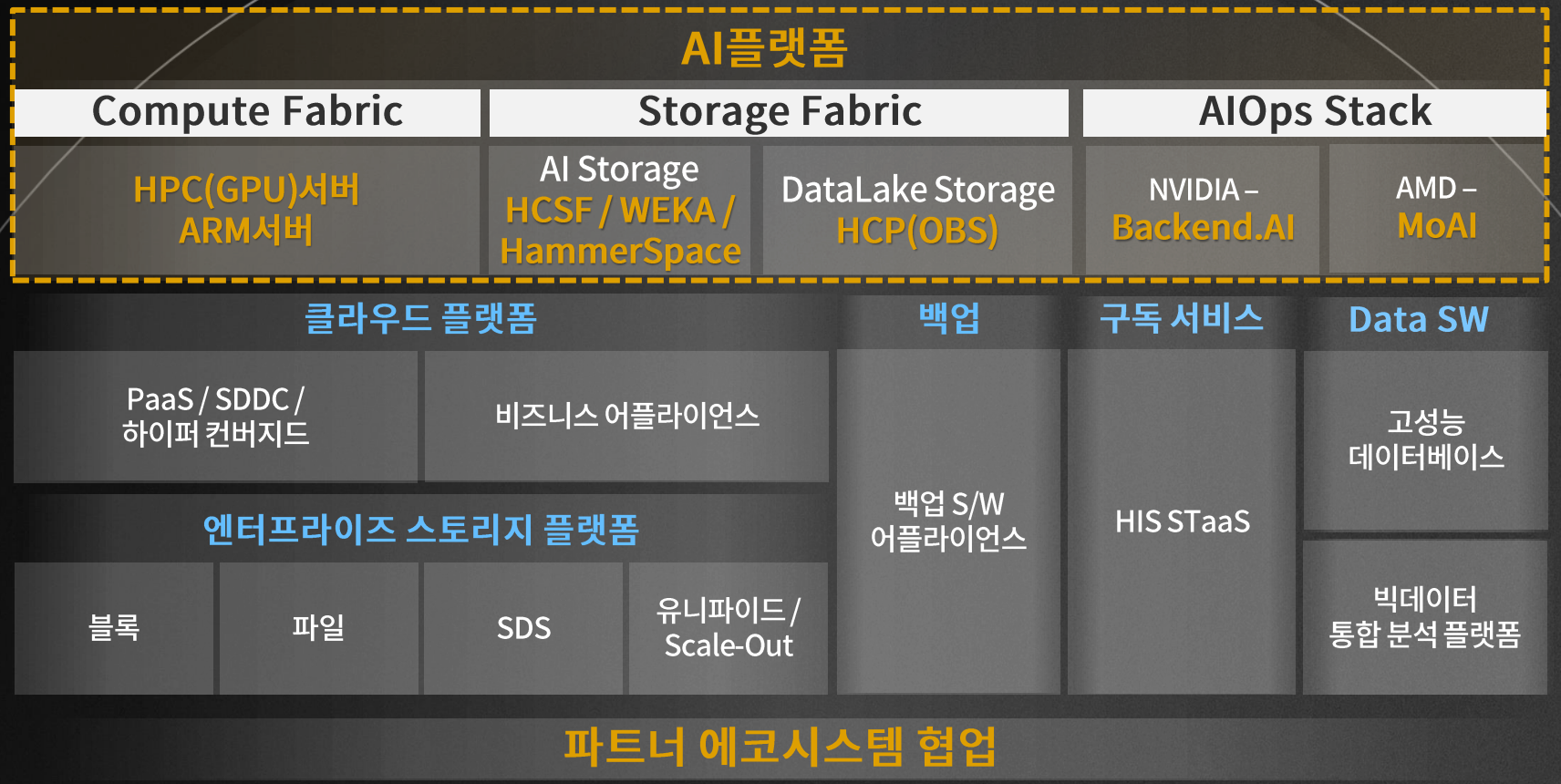
II. HS효성 AI플랫폼

1. Compute Fabric
2. Storage Fabric
3. AIOps Stack
4. 저전력 Arm서버 – GreenCore
5. 데이터통합 - HammerSpace



HS효성 AI플랫폼

AI인프라를 위한 시너지 강화



Compute Fabric - Supermicro

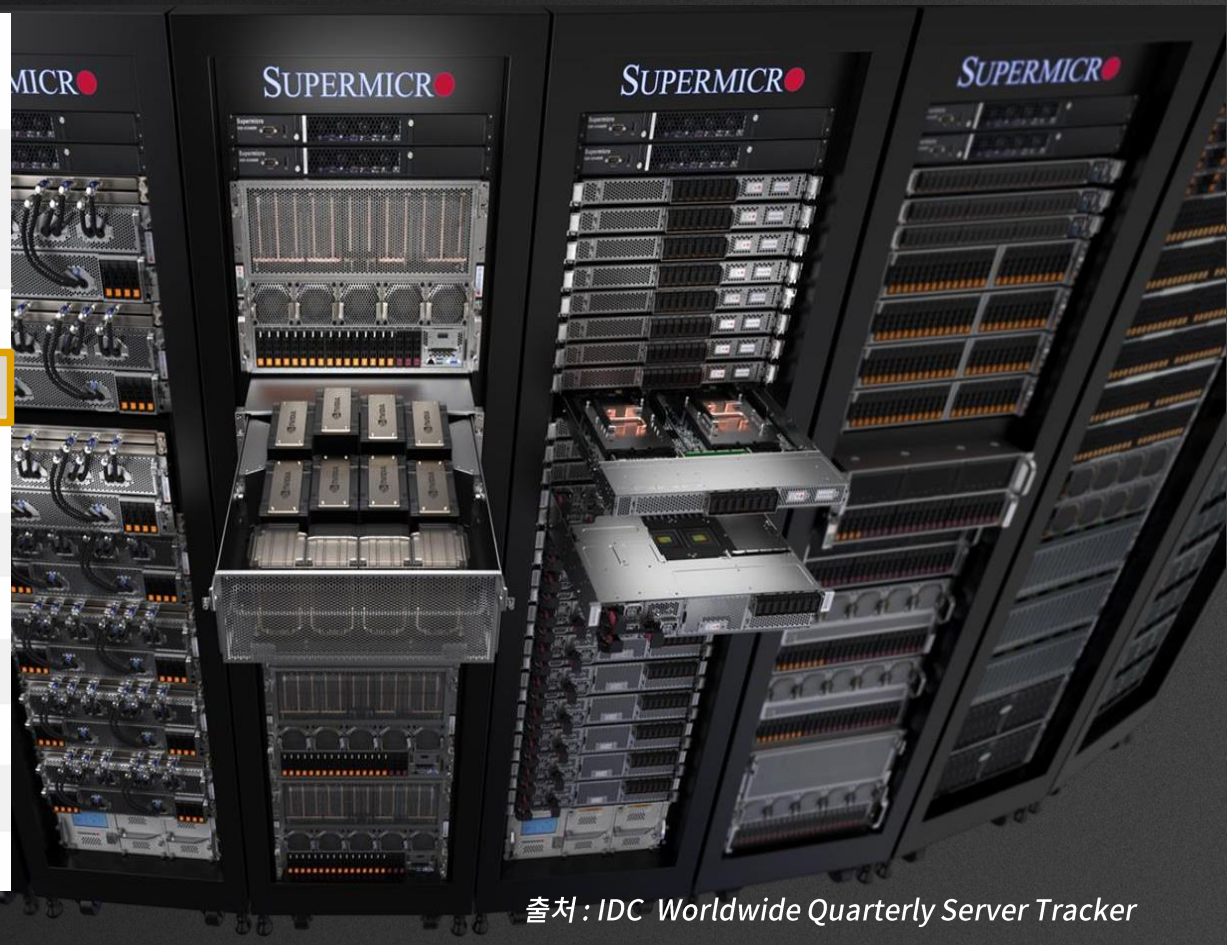
- 글로벌 GPU 서버 시장 1위 (전년 대비 55% 성장)
- FTM 전략(First to Market)을 통해 빠른 개발 및 품질 유지 → 출시 과정을 간소화

Top 5 Companies, Worldwide Server Market, Fourth Quarter of 2024

(Vendor Revenue in US$ millions)

Company	4Q24 Vendor Revenue	4Q24 Market Share	4Q23 Vendor Revenue	4Q23 Market Share	4Q24/4Q23 Revenue Growth
T1. Dell Technologies*	$5,540.4	7.2%	$4,592.6	11.3%	20.6%
T1. Super Micro*	$5,005.1	6.5%	$3,230.1	8.0%	55.0%
T2. Hewlett Packard Enterprise*	$4,239.1	5.5%	$2,749.2	6.8%	54.2%
T2. IEIT Systems*	$3,878.1	5.0%	$2,333.6	5.8%	66.2%
T2. Lenovo*	$3,783.7	4.9%	$2,226.1	5.5%	70.0%
ODM Direct	$36,570.2	47.3%	$14,313.1	35.4%	155.5%
Rest of Market	$18,304.6	23.7%	$11,035.9	27.3%	65.9%
Total	**$77,321.1**	**100.0%**	**$40,480.6**	**100.0%**	**91.0%**

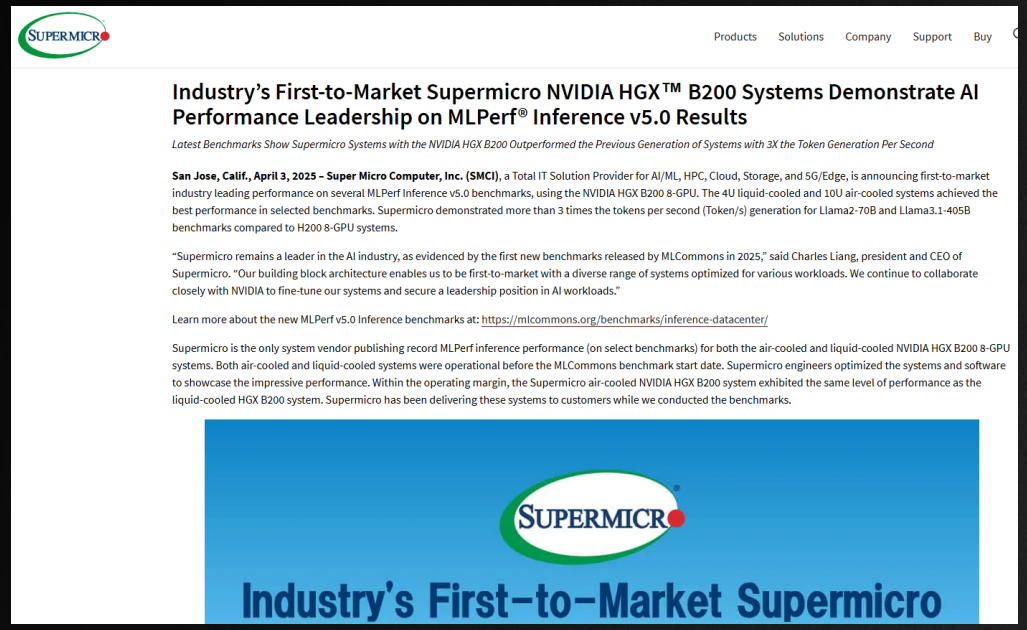
Source: IDC Worldwide Quarterly Server Tracker, March 13, 2025.



출처 : IDC Worldwide Quarterly Server Tracker

Compute Fabric - Supermicro_FTM 전략

- 공랭식 및 수랭식 NVIDIA HGX B200 8-GPU 시스템 모두에서 MLPerf 추론 성능 신기록을 달성한 유일한 시스템 공급업체
  - Llama2-70B 및 Llama3.1-405B 벤치마크에서 H200 8-GPU 시스템 대비 3배 이상의 초당 토큰(Token/s) 생성을 입증 (일부 벤치마크 기준)



SUPERMICRO

Products Solutions Company Support Buy

Industry's First-to-Market Supermicro NVIDIA HGX™ B200 Systems Demonstrate AI Performance Leadership on MLPerf® Inference v5.0 Results

Latest Benchmarks Show Supermicro Systems with the NVIDIA HGX B200 Outperformed the Previous Generation of Systems with 3X the Token Generation Per Second

San Jose, Calif., April 3, 2025 – Super Micro Computer, Inc. (SMCI), a Total IT Solution Provider for AI/ML, HPC, Cloud, Storage, and 5G/Edge, is announcing first-to-market industry leading performance on several MLPerf Inference v5.0 benchmarks, using the NVIDIA HGX B200 8-GPU. The 4U liquid-cooled and 10U air-cooled systems achieved the best performance in selected benchmarks. Supermicro demonstrated more than 3 times the tokens per second (Token/s) generation for Llama2-70B and Llama3.1-405B benchmarks compared to H200 8-GPU systems.

"Supermicro remains a leader in the AI industry, as evidenced by the first new benchmarks released by MLCommons in 2025," said Charles Liang, president and CEO of Supermicro. "Our building block architecture enables us to be first-to-market with a diverse range of systems optimized for various workloads. We continue to collaborate closely with NVIDIA to fine-tune our systems and secure a leadership position in AI workloads."

Learn more about the new MLPerf v5.0 Inference benchmarks at: https://mlcommons.org/benchmarks/inference-datacenter/

Supermicro is the only system vendor publishing record MLPerf inference performance (on select benchmarks) for both the air-cooled and liquid-cooled NVIDIA HGX B200 8-GPU systems. Both air-cooled and liquid-cooled systems were operational before the MLCommons benchmark start date. Supermicro engineers optimized the systems and software to showcase the impressive performance. Within the operating margin, the Supermicro air-cooled NVIDIA HGX B200 system exhibited the same level of performance as the liquid-cooled HGX B200 system. Supermicro has been delivering these systems to customers while we conducted the benchmarks.

MLPerf Results

MLCommons data as of: 3/25/2025 10:49:49 PM

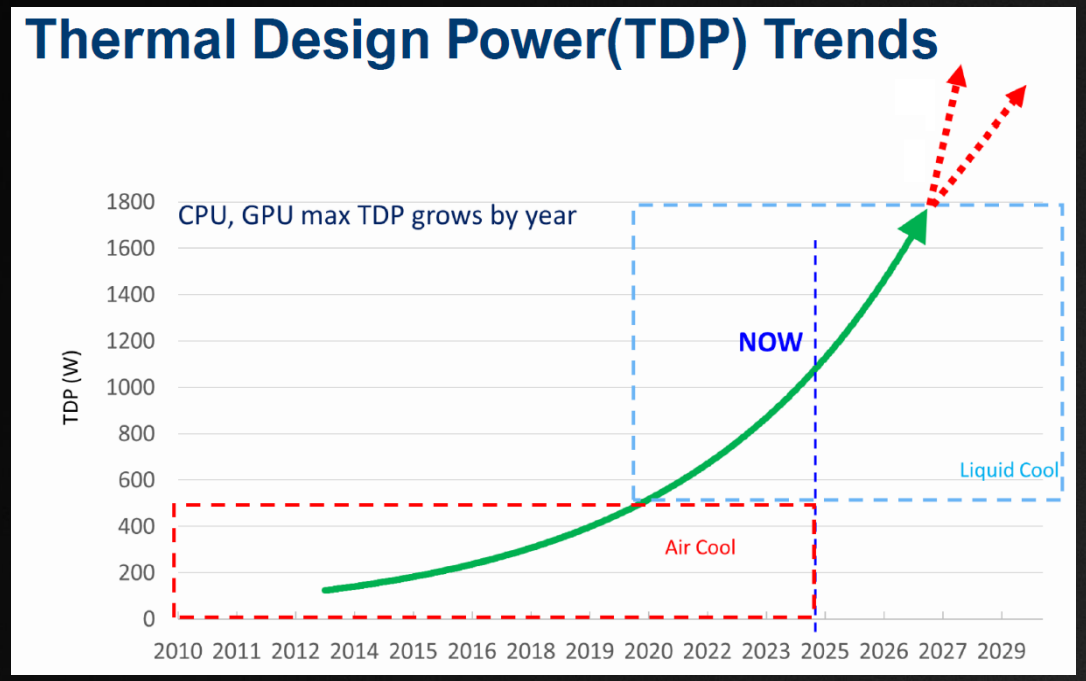
Benchmark: **Inference**
System Type: **Datacenter**
Division/Power: **Closed**
Availability: **available**
Round: **v5.0**

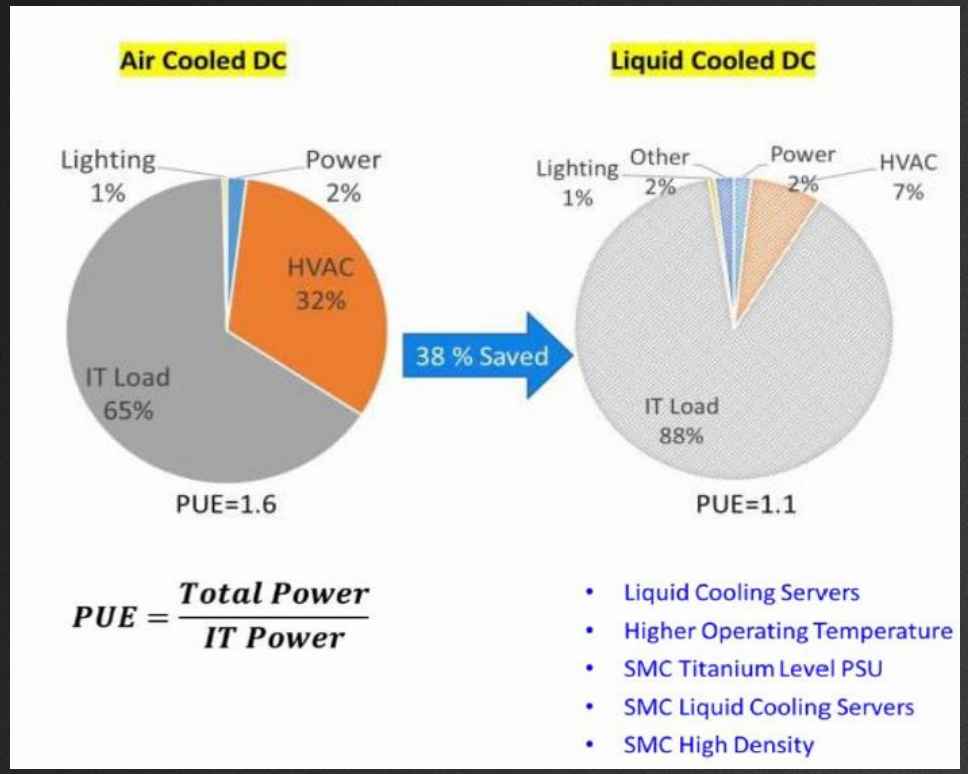
							Benchmark / Model MLC / Scenario / Units			
							1-405b		mixtral-8x7b	re
							Server	Offline	Server	Offline
Public ID	Organization	Availability	System Name (click + for details)	# of Nodes	Processor	Accelerator	# of Accelerators	Queries/s	Tokens/s	Tokens/s	Samples/s
5.0-0067	Supermicro	available	SYS-822GA-NGR3	1	INTEL(R) XEON(R) 6980P	N/A	0				46,041.40
5.0-0068	Supermicro	available	AS-8126GS-TNMR	1	AMD EPYC 9575F	AMD Instinct MI325X 25..	8				
5.0-0069	Supermicro	available	ARS-111GL-NHR (1x GH200-96GB_aarch..	1	NVIDIA Grace CPU	NVIDIA GH200 Grace H..	1		7,718.34	7,182.77	
5.0-0070	Supermicro	available	AS-4125GS-TNHR2-LCC (8x H200-SXM-..	1	AMD EPYC 9654	NVIDIA H200-SXM-141..	8	290.09	62,766.20	59,647.10	
5.0-0071	Supermicro	available	SYS-522GA-NRT(8x H200-NVL-141GB)	1	Intel(R) Xeon(R) 6979P	NVIDIA H200-NVL-141..	8				
5.0-0072	Supermicro	available	SYS-421GE-NBRT-LCC (8x B200-SXM-1..	1	Intel(R) Xeon(R) Platinum 85..	NVIDIA B200-SXM-180..	8	1,057.52	128,795.00	129,047.00	
5.0-0073	Supermicro	available	SYS-821GE-TNHR (8x H200-SXM-141GB..	1	Intel(R) Xeon(R) Platinum 85..	NVIDIA H200-SXM-141..	8		61,964.10	59,627.30	759,826.00
5.0-0074	Supermicro	available	SYS-A21GE-NBRT (8x B200-SXM-180GB..	1	Intel(R) Xeon(R) Platinum 85..	NVIDIA B200-SXM-180..	8	1,080.31	125,534.00	128,961.00	

출처 : https://mlcommons.org/benchmarks/inference-datacenter/

Compute Fabric - Supermicro_Liquid Cooling

- AI데이터센터의 전력 효율화(PUE < 1.1)를 위한 냉각 방식 – **Liquid Cooling**
- 효율적인 열교환 방식의 빠른 냉각을 통한 **AI work-flow 전반 성능 향상!**





*PUE: Power Usage Effectiveness, 데이터 센터의 에너지 효율성을 나타내는 대표적인 지표

Compute Fabric - Supermicro_DLC-2

- **Supermicro DLC-2**: 고성능과 에너지 효율을 동시에 실현하는 **차세대 직접 수냉식 솔루션**
- 서버, CDU 등 부품 일체를 Supermicro **직접 생산**

구분	DLC (1.0)	DLC 2.0 (신규)
PUE	1.05 (목표)	
냉각 커버리지	CPU, GPU	CPU, GPU, Memory, PSU, VRM, PCIe switch등 (약 98%)
유입 온도	30~35°C	최대 45°C
CDU 용량	랙당 100kW	랙당 250kW
소음	약 73dB	약 50dB
서버 밀도	표준	약 60% 상면 개선
냉각 시스템 관리	부분적 통합	SuperCloud Composer 완전 통합 관리
전력 절감	10~20% 전력 절감	최대 40% 전력 절감
냉각용수 절감	-	최대 40% 액체 절감

*CDU: Coolant Distribution Unit, 수냉식 시스템에서 냉각수를 관리하고 분배하는 핵심 장치

Storage Fabric - HCSF / WEKA (1/2)

- AI 업무를 위해 필요한 IOPS 성능, 다양한 인터페이스, 대용량 확장을 제공하는 고성능 스토리지

초고성능 분산 병렬 파일시스템(NVMe) (WekaFS)

- ✓ 최신 기술 적용으로 로컬 플래시 드라이브 보다 2~3배 이상 빠른 초고성능 보장 / 파일 크기 무관
- ✓ 다양한 고속화 프로토콜 지원 (DPDK*, GDS* 등)
- ✓ 메타데이터 분산 처리를 통한 병목 항목 제거

대용량 Object Storage (HCP)

- ✓ 무제한에 가까운 확장성
  - 비용 효율적인 대용량 데이터 관리
  - 정책 기반의 티어링 운영
- ✓ 선형적 성능 및 용량 확장
- ✓ 안정적인 티어링 성능

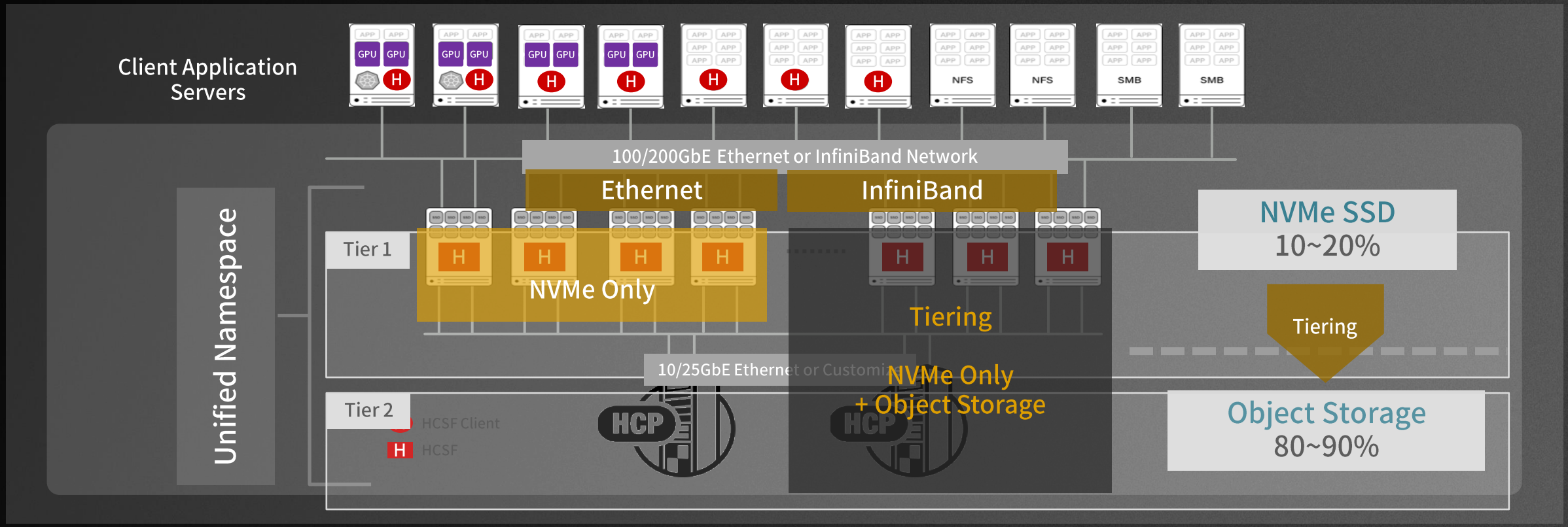
HCSF

- ✓ 안정적인 데이터 처리 환경 제공
- ✓ 파일 크기 무관 모두 고성능 제공
- ✓ 다양한 프로토콜 호환성으로 유연한 어플리케이션 연동 환경 지원
- ✓ 이상적인 데이터 분석용 스토리지

*DPDK: Data Plane Development Kit 의 약어로써, 네트워크 패킷 고속 처리 기술
*GDS: NVIDIA GPU Direct Storage 의 약어. GPU – Storage 병목 구간 고속 처리 기술

Storage Fabric - HCSF / WEKA (2/2)

- HCSF: 고성능 분산 병렬 파일시스템 어플라이언스
- Tier1: HPC 환경에 적합한 All NVMe 기반의 초고성능 분산 병렬 파일시스템
- Tier2: 대규모 Cold Data를 안전하고 비용 효율적으로 티어링(정책기반, 자동화)해 보관 및 관리할 수 있는 오브젝트 스토리지

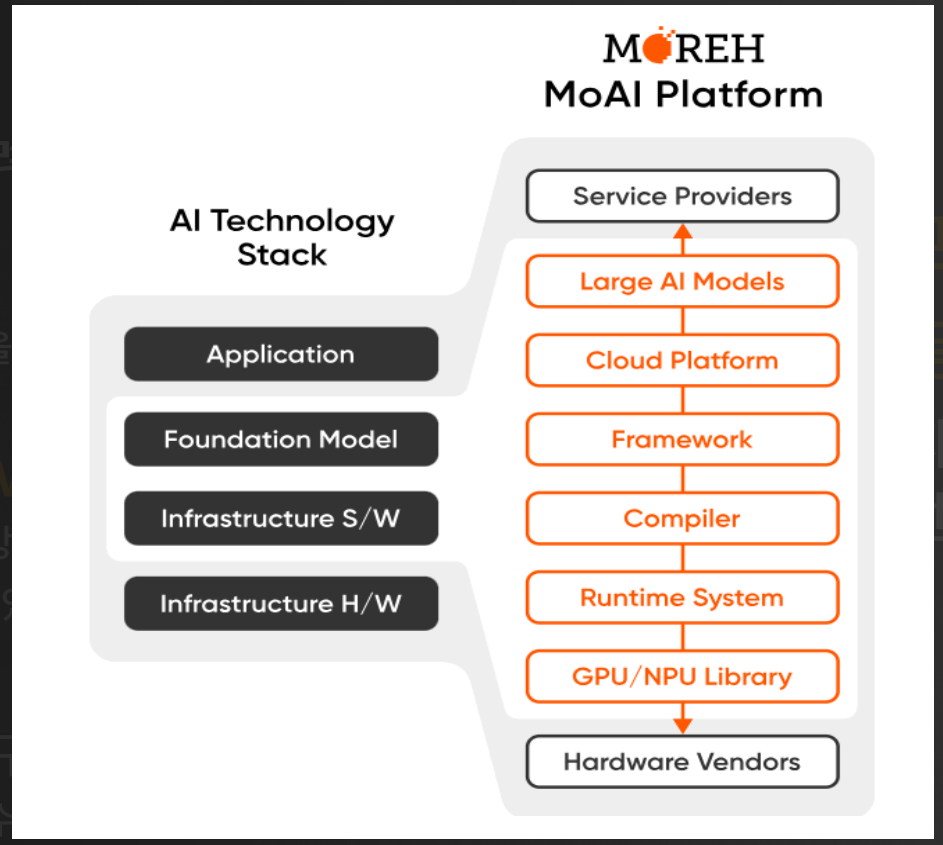


- 티어링 시 구성 용량은 요구 성능 및 환경에 따라 달라질 수 있음

AIOps Stack - Backend.AI / MoAI

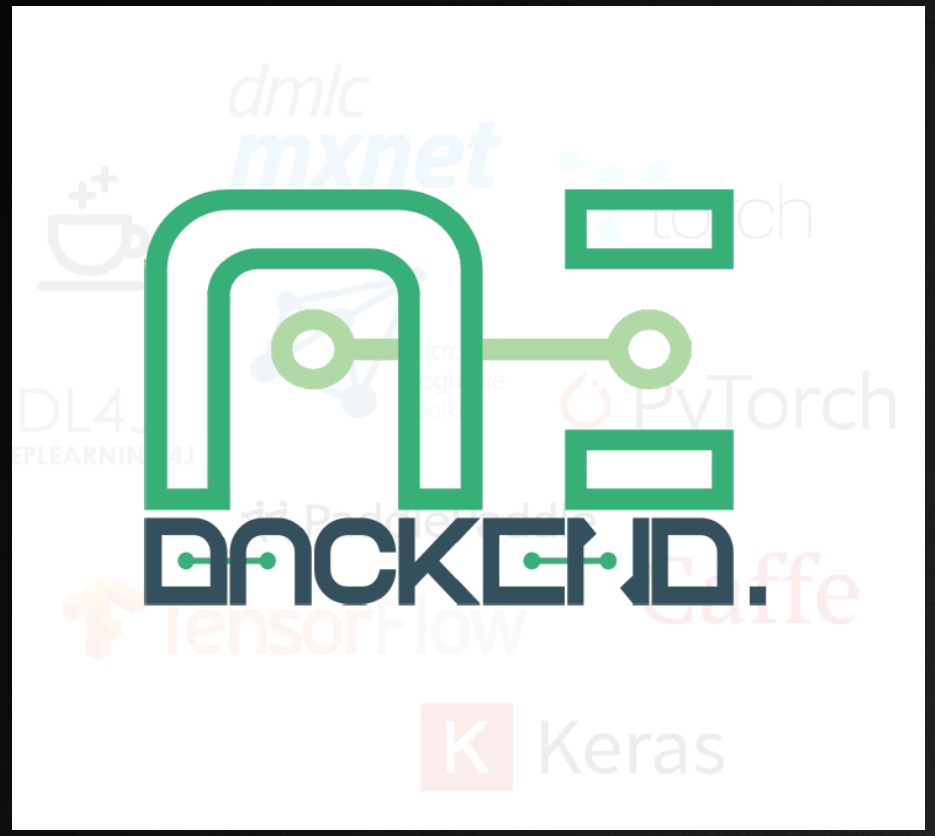
NVIDIA
NVIDIA DGX-Ready Software 국내 기업

AMD
AMD 시리즈 B 투자 - 국내 기업



AIOps Stack - Backend.AI (1/2)

- 아태지역 최초의 NVIDIA DGX-Ready Software
- 검증된 AI 플랫폼으로 국내 선도 대기업 대상 다수의 사례 보유



1 GPU 활용 극대화

- 컨테이너 수준 GPU 분할 가상화
- 정책 기반 자원 관리

2 직관적인 사용자 환경

- GUI 기반 컨테이너 운영관리
- 웹UI와 데스크탑 앱 지원

3 사전정의 AI 개발 환경 제공

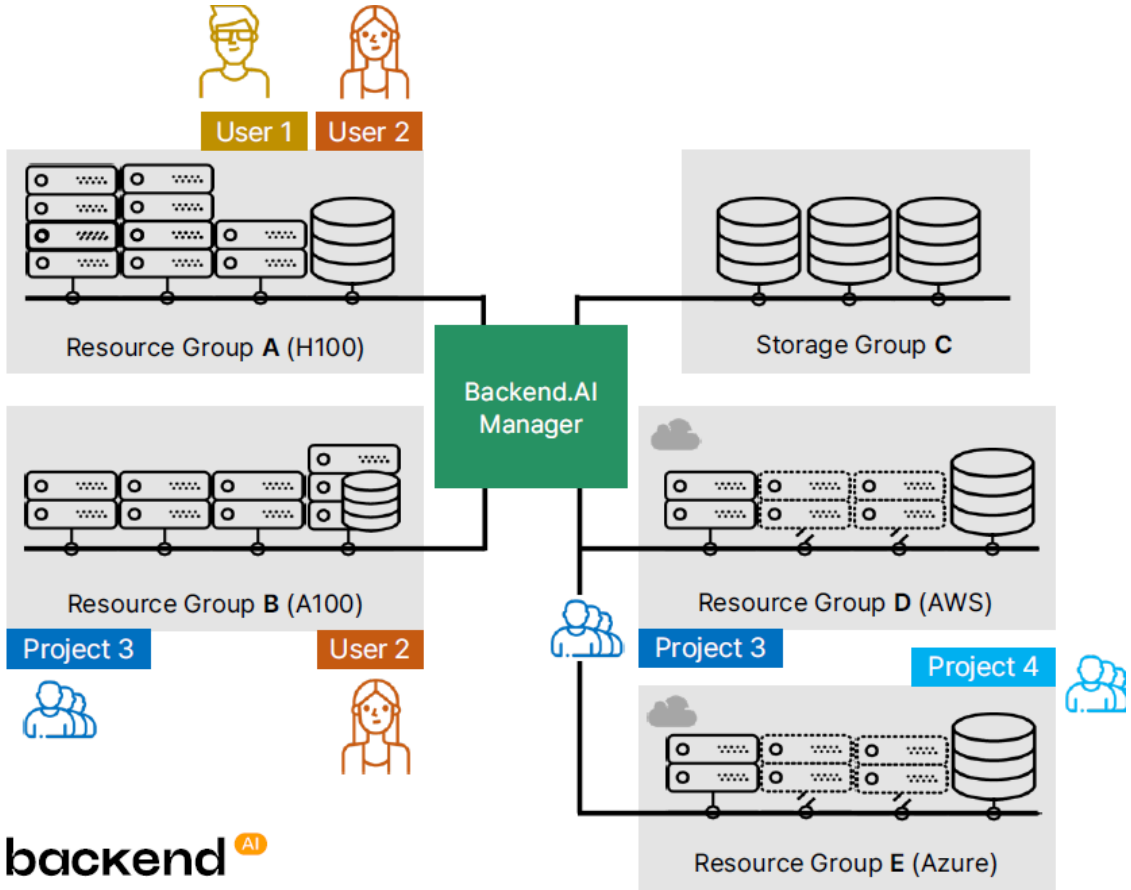
- Tensorflow, Pytorch 등 사전정의 이미지 제공
- 연구 환경 선택 즉시 생성

4 AI 및 HPC 성능 최적화

- 고성능 병렬 파일 스토리지와 성능 최적화 구현
- 독자적 엔진으로 최적의 GPU 연산 자원 배치 구현

AIOps Stack - Backend.AI (2/2)

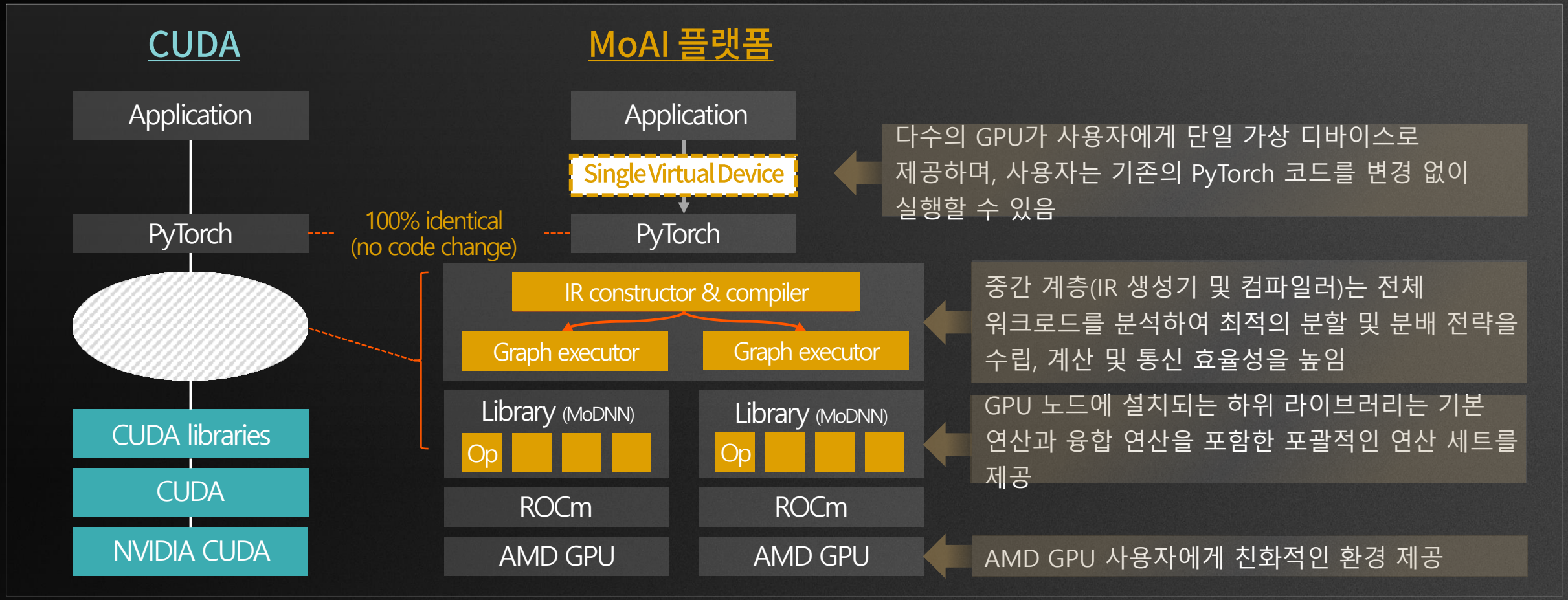
- 아태지역 최초의 NVIDIA DGX-Ready Software
- 검증된 AI 플랫폼으로 국내 선도 대기업 대상 다수의 사례 보유



- 다른 성능의 GPU를 한 번에 효율적으로 관리
- 팀별/부서별/사용자별 완벽한 자원 격리
- GPU Utilization 극대화

AIOps Stack - MoAI (1/2)

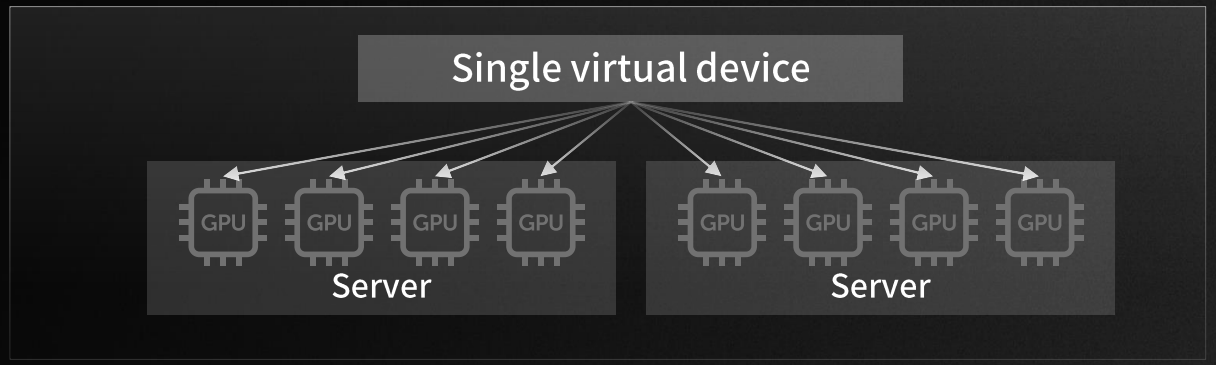
- AMD GPU + MoAI 솔루션으로 비용효율적 구성 가능
- CUDA Library 종속성 탈피



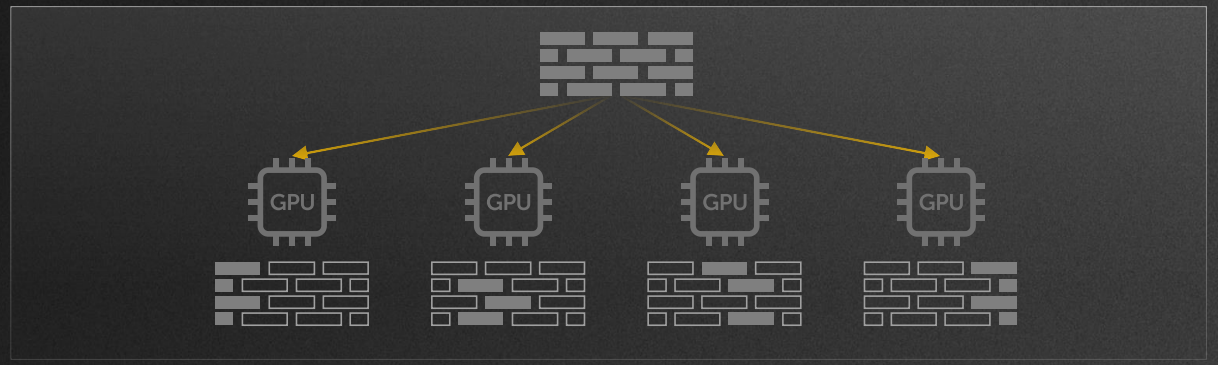
AIOps Stack - MoAI (2/2)

- 개발편의성, Network 비용 절감, HW 장애 대응

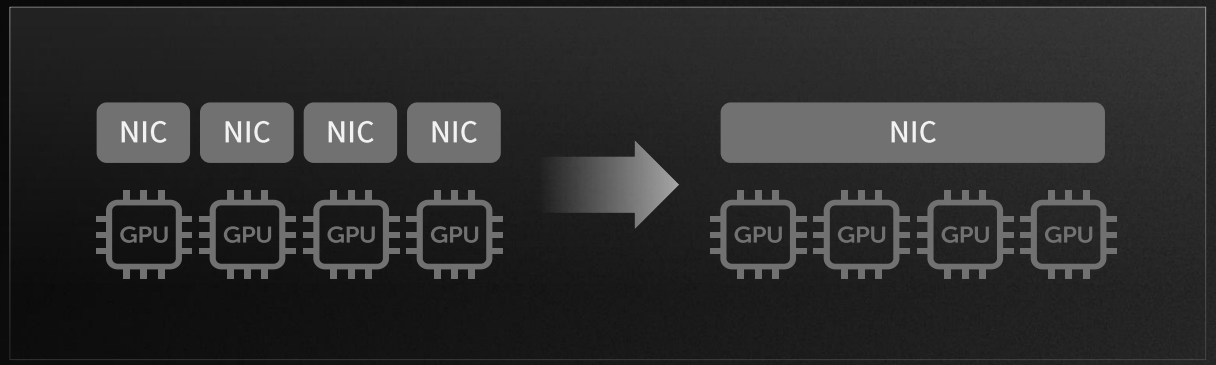
N-1 & 1-N GPU 가상화



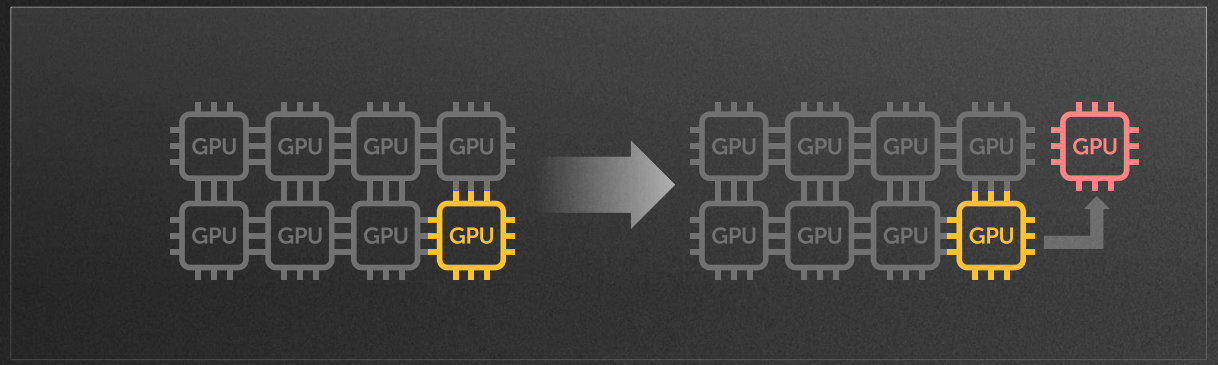
자동 병렬화



Network 최적화



Fault Tolerance



저전력 Arm서버

저전력 & 저발열 서버

- 모바일 최적화된 아키텍처 → 저전력
- 칩 자체 발열 Low
  → 냉각 비용 절감

도입원가 절감

- X86 대비 비용 효율성 우수
- 주요 부품 수량 감소로 전체 HW 비용 절감

전력소모 대비 뛰어난 성능

- 단일 칩 아키텍처로 일관된 성능 제공

aws

Google Cloud

Microsoft Azure

NVIDIA

저전력 Arm서버 - GreenCore

- **GreenCore**: HS효성인포메이션시스템 & 엑세스랩 공동 개발한 **국산 Arm서버**

GreenCore

H/W 보드 직접 설계, 개발

국내 유일 Arm 보드 일체 설계/개발

S/W 일체형 지원

Arm 관련 OS 및 다양한 오픈소스에 대한 가이드 지원

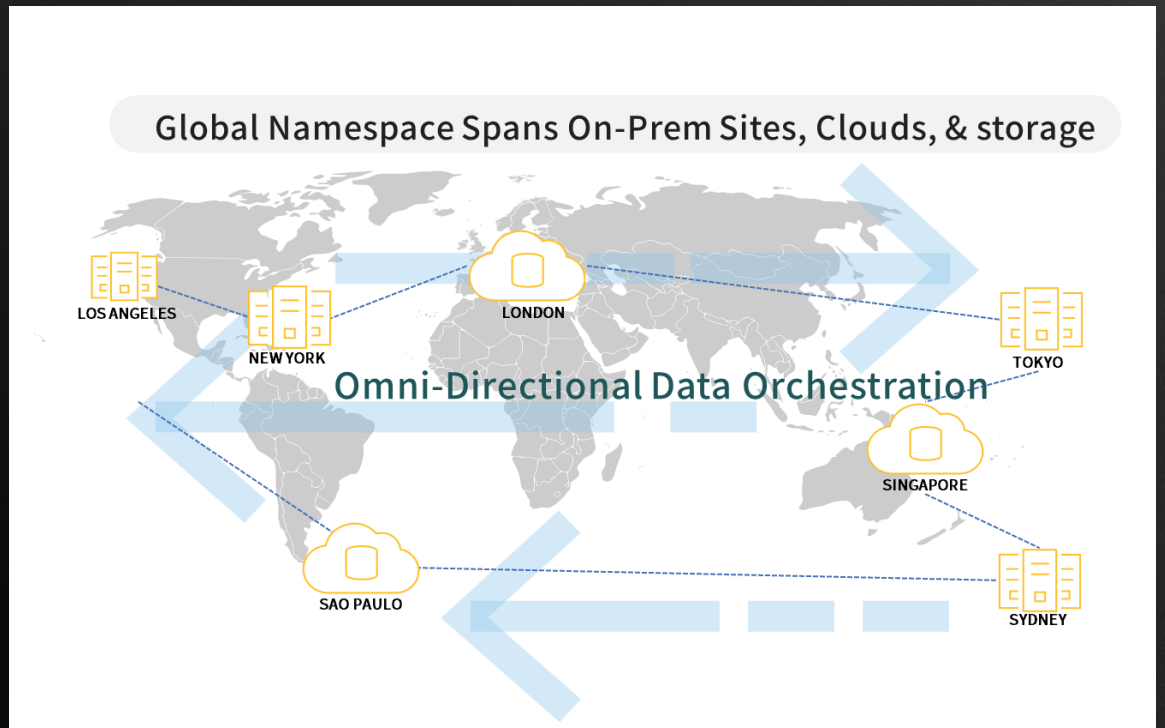
맞춤형 관리환경

Arm 서버 전용 BMC 제공 및 관리 GPU 화면 제공

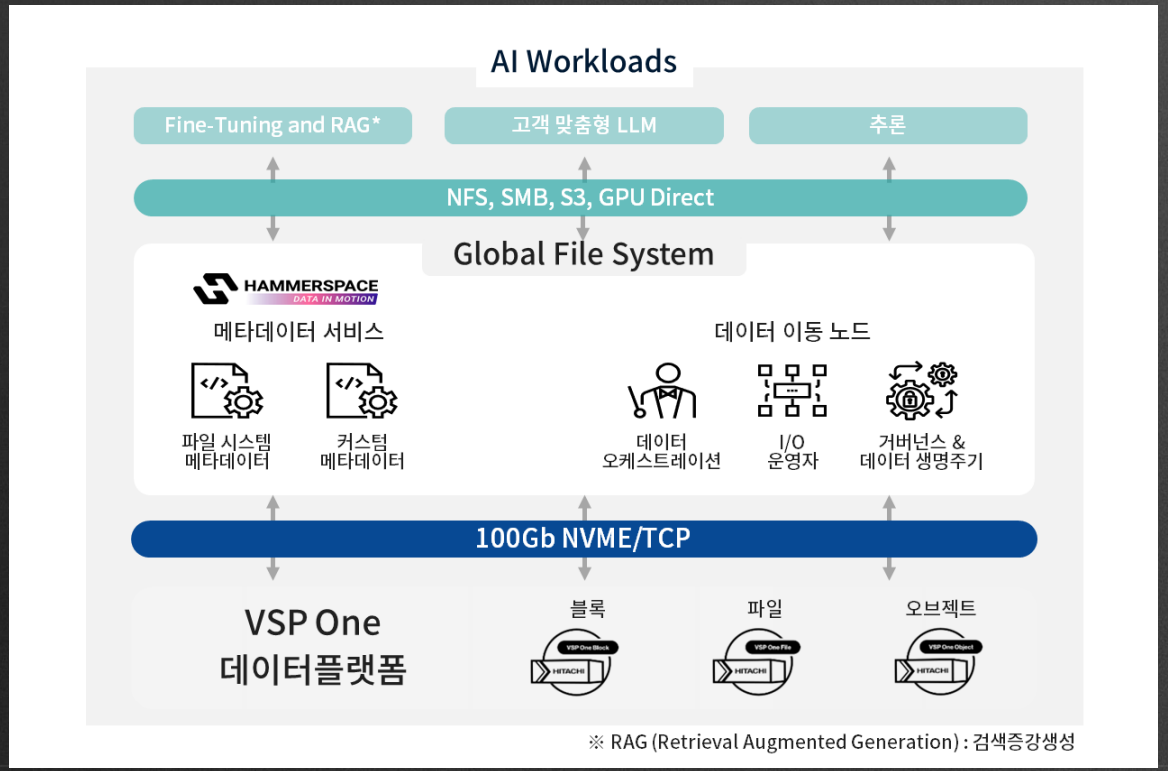
데이터통합 - HammerSpace (1/2)

- AI 시대, 더 중요해진 데이터 관리! – 데이터 활용도 극대화 및 데이터 사일로 해소 필요!
- 글로벌파일시스템 해머스페이스를 통한 데이터 통합 관리

데이터 오케스트레이션: 위치의 제약 없는 데이터 운영



글로벌파일시스템: 데이터 통합 관리

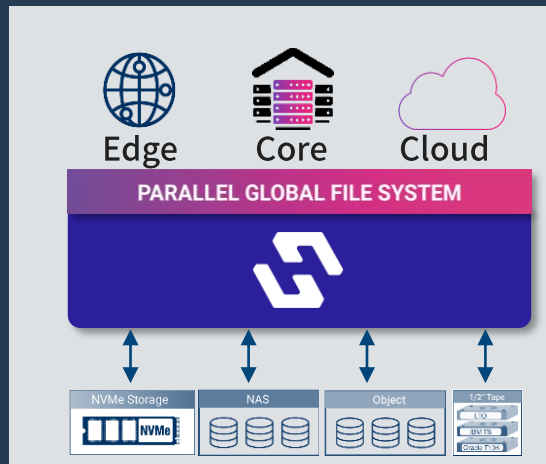


※ RAG (Retrieval Augmented Generation) : 검색증강생성

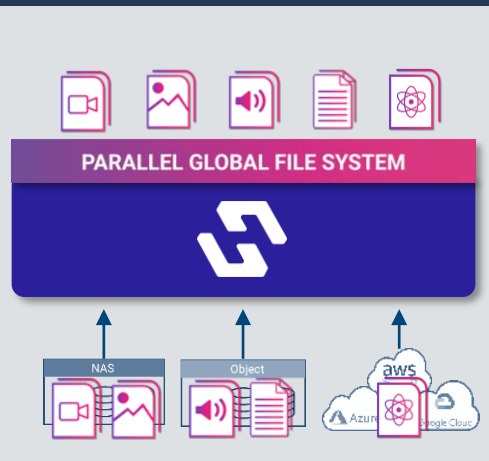
데이터통합 - HammerSpace (2/2)

- AI 시대, 더 중요해진 데이터 관리! – 데이터 활용도 극대화 및 데이터 사일로 해소 필요!
- 글로벌파일시스템 해머스페이스를 통한 데이터 통합 관리

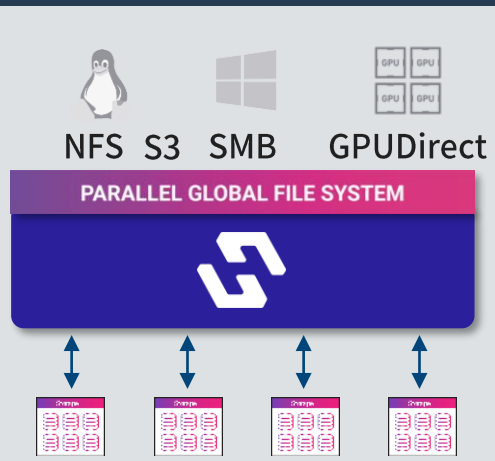
모든 환경, 모든 스토리지 구축



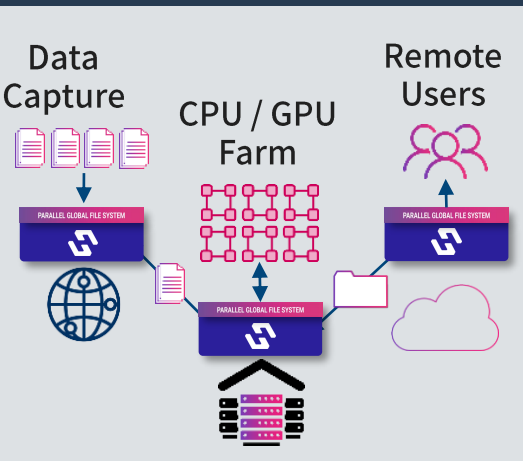
물리적 데이터 이관 X



병렬 처리를 위한 빠른 데이터 액세스 제공

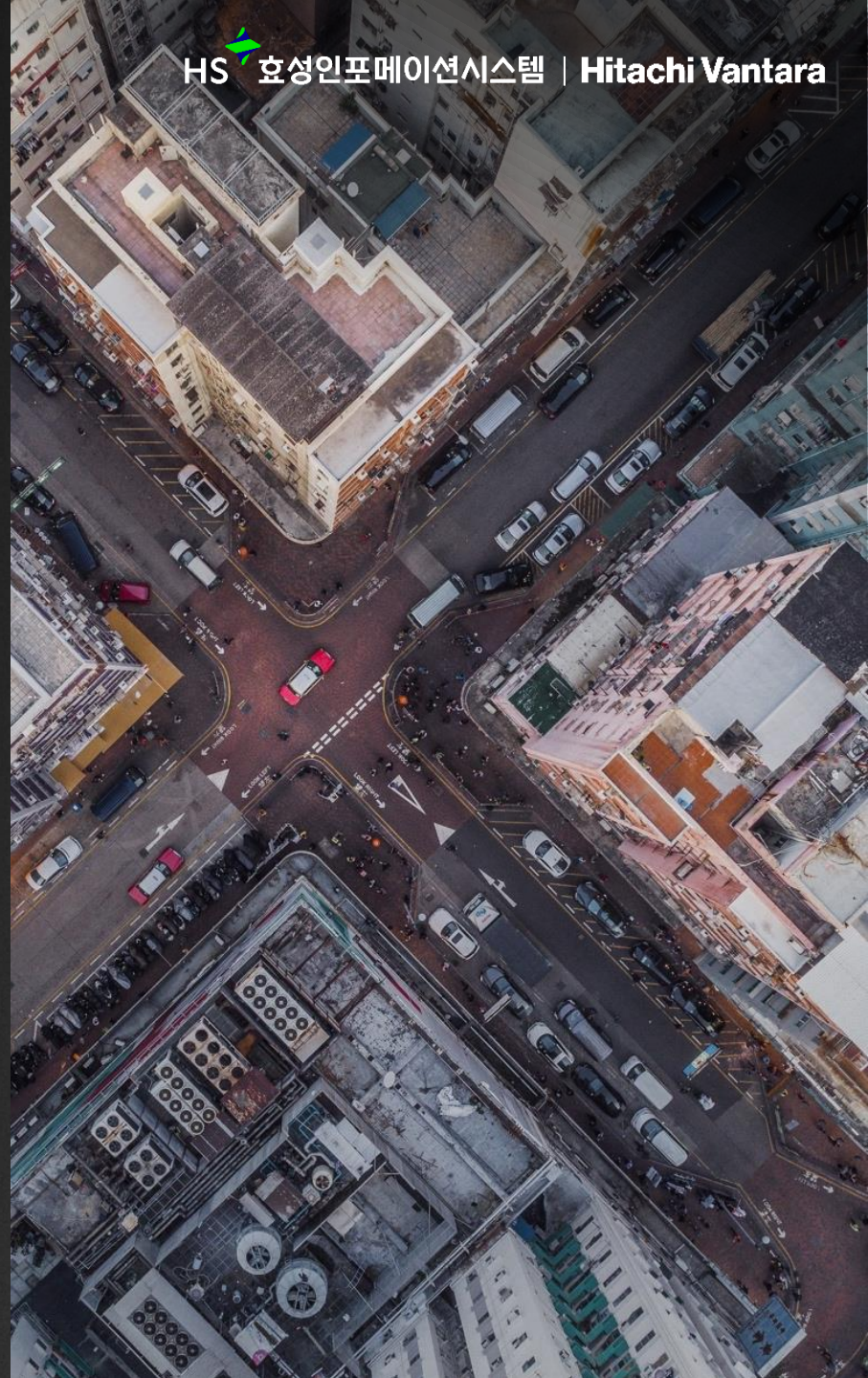


정책에 따른 손쉬운 데이터 자동 배치



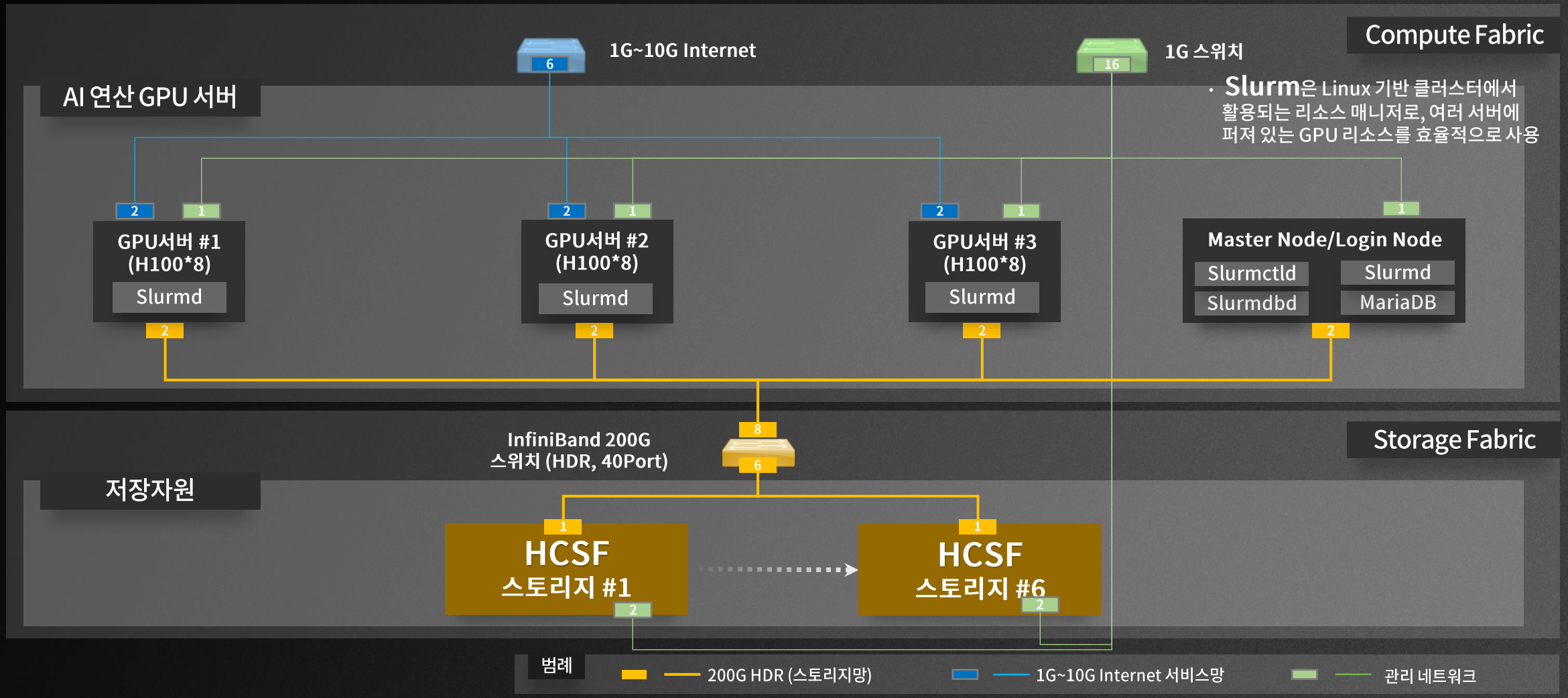
III. 대표 구축 사례

1. IT 대기업 AI플랫폼 인프라(자체 LLM 개발)
2. 대기업 DX GPU AI 인프라 구축(연구 개발)



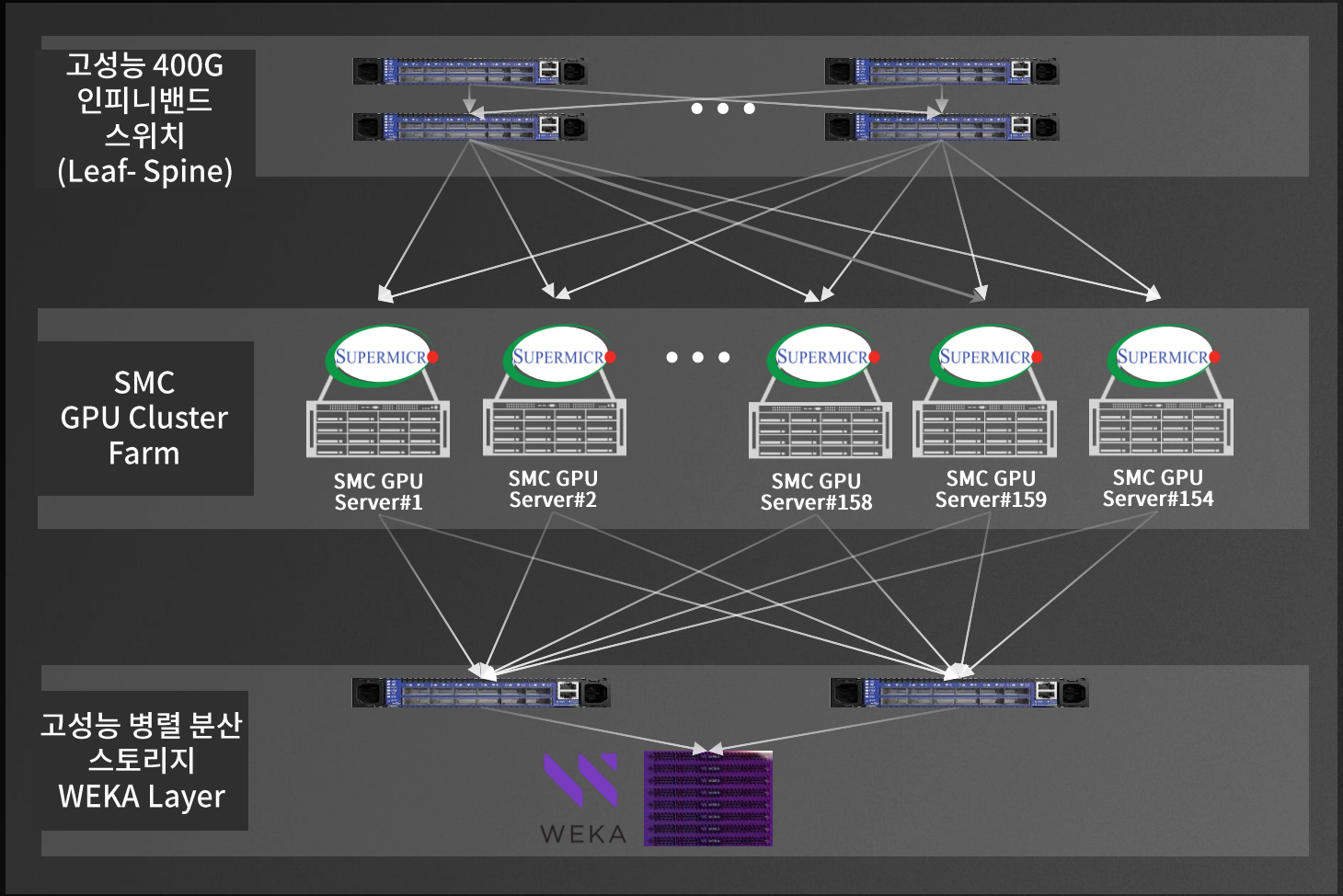
IT 대기업 AI플랫폼 인프라(자체 LLM 개발)

- HPC 클러스터(HW & SW) & HCSF(분석특화 저장소) & 오픈소스 클러스터 관리 솔루션 구축 사례



대기업 DX GPU AI 인프라 구축(연구 개발)

- **고성능 AMD GPU Server Cluster Farm 구축 사례**



사업 목적

- **고객사 DX GPU AI 인프라 구축** 목적의 고성능 **AMD GPU Cluster Farm** 인프라 도입 및 구성
- 운영 **154** GPU Cluster Farm / 개발 **6** GPU Cluster Farm 구축: 총합 160대 도입

SuperMicro + HIS 선정 이유

- SuperMicro Server의 모듈식 설계에 따른 **유연성**과 **확장성**, 서버 성능을 보장하는 **안정성**
- HIS의 전문 기술지원 인력을 통한 최고의 **직접 기술지원 서비스**와 **HPC/고성능 스토리지/AI 구축 레퍼런스**

AMD GPU 선정 이유

- One Vendor GPU(NVIDIA) **종속성 탈피** 및 **비용절감**을 위한 고성능 AMD GPU가 장착된 안정적인 **고성능 SMC GPU Server** 도입

AI 인프라 도입 시 고려 사항

맺음말

다양한 에코 파트너 협업 체계 구축 확인

빠르게 변화하는 AI시대 대응

AI플랫폼 구축 경험 확인

Compute Fabric, Storage Fabric, AIOps Stack

국내외 실 사례를 통한 국내 기술력(인력) 여부 확인

장애 지원, 신규 AI 솔루션 연계 지원

감사합니다.