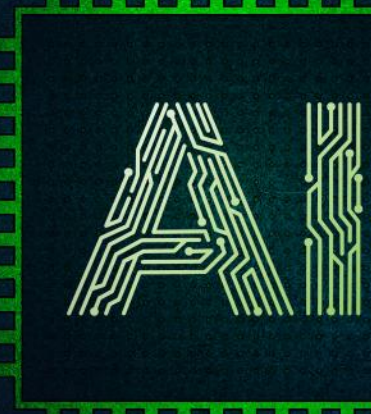
HS 효성인포메이션시스템 | HITACHI

AI 인프라슈퍼사이클

엔터프라이즈 AI 멀티에이전트 운영을 위한 AI 인프라 구축 전략



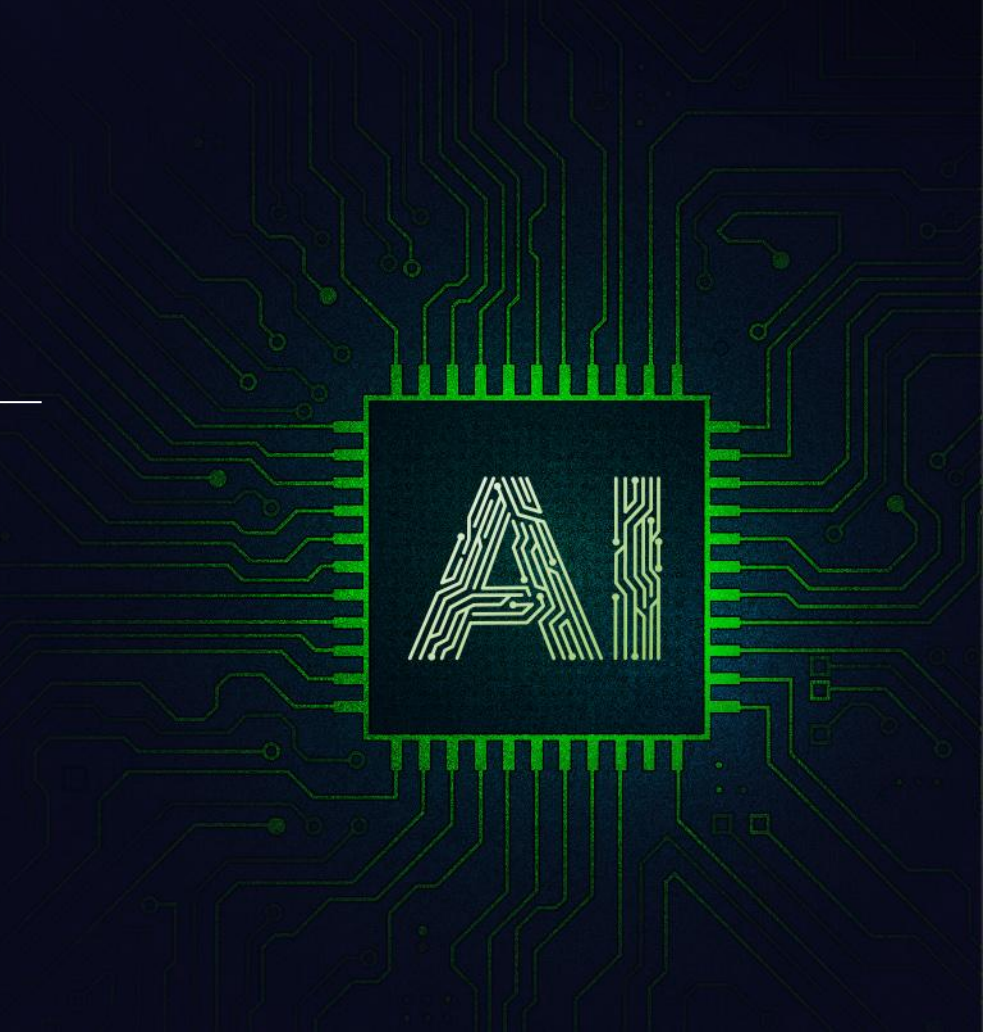
DX아키텍트팀 권동수 전문위원(his-dskwon@hshyosung.com)

Agenda

1. AI 트렌드
2. 멀티에이전트 구축을 위한 AI플랫폼
3. AI 구축 방안 및 사례

1.
AI 트렌드

1. 지속가능한 미래를 위한 AI
2. 멀티에이전트 AI 인프라 4대 필수 요건
3. AI 기술 발전 전망
4. AI 인프라 구축의 주요 구성 요소



1. 지속가능한 미래를 위한 AI

멀티모달 AI

텍스트, 이미지, 음성 등
다양한 데이터를 동시에 처리

Agentic AI

개인화된 AI Agent를 통한
단순 작업 자동화에서 **다중 단계 업무까지 수행**

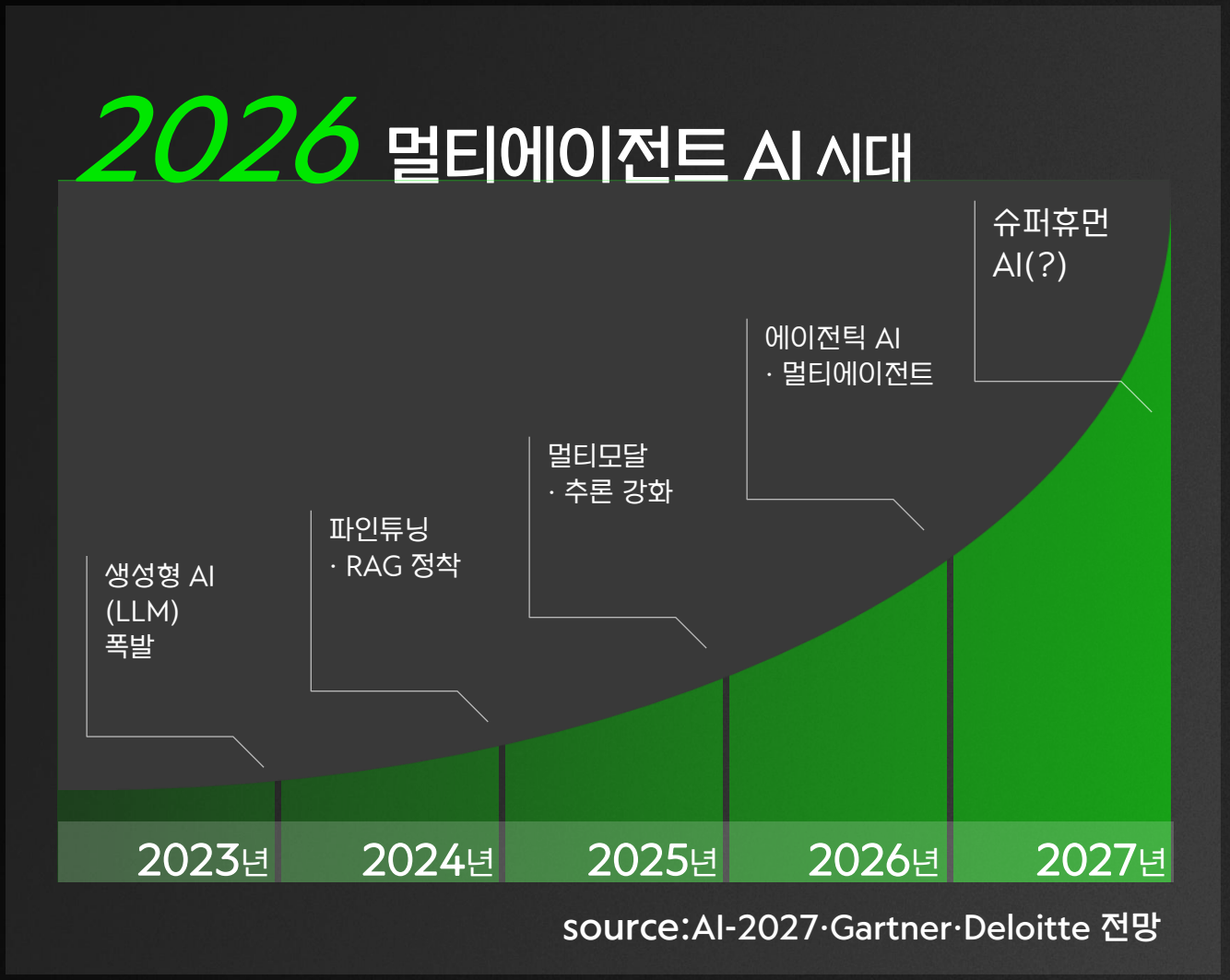
Physical AI

물리적 법칙 + 데이터 기반
학습을 통해 실제 현상을
보다 정확히 예측

2. 멀티에이전트 AI 인프라 4대 필수 요건

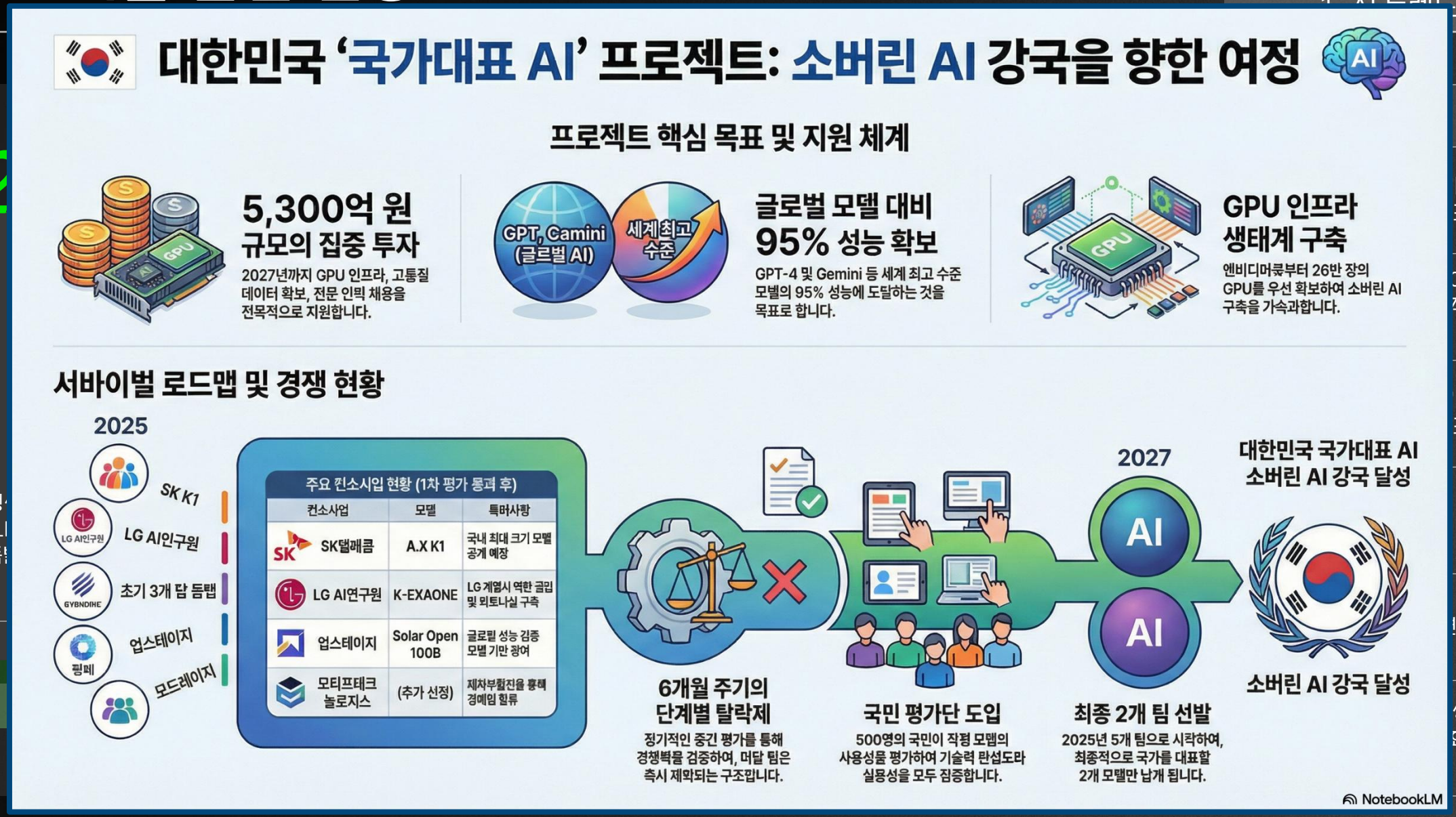
01 저지연 추론 및 가용성	◦ 에이전틱 AI는 목표 달성을 위해 '추론-행동-피드백'의 반복을 거치므로, 저지연 실행 환경이 비즈니스 성패를 결정하는 필수 요건
02 통합 메모리 및 데이터 기반	◦ 에이전트가 스스로 판단하려면 기업 내 분산된 데이터를 실시간으로 읽고 맥락을 기억하는 'AI 전용 데이터 플랫폼'이 필요
03 연결성 및 도구 오케스트레이션	◦ 에이전트가 단독으로 존재하지 않고 시스템 간 협업(Collaborative Agents)을 해야 하므로, 상호운용성 표준(Interoperability standards)과 API 연결 인프라가 필수
04 통합 모니터링 및 거버넌스	◦ 에이전틱 MLOps (AgentOps): 에이전트의 추론 단계별 로그를 기록하고, 오류 발생 시 즉시 개입할 수 있는 모니터링 시스템이 필요

3. AI 기술 발전 전망



모델명	특징 요약
Gemma 4 2026년 4월 발표	• 아파치 2.0 라이선스로 제공, 전문가 혼합(MoE) 구조의 AI 모델(양자화 포맷 지원)
Llama 4.5 (Enhanced) 2026년 1월 발표	• 10M 토큰의 초거대 컨텍스트와 네이티브 멀티모달 능력 및 에이전트 성능 최적화 (양자화 포맷 지원)
gpt-oss-120B 2025년 8월 발표	• 강력한 추론 성능을 제공하는 전문가 혼합(MoE) 구조의 AI 모델(양자화 포맷 지원)
Llama 4 Maverick 2025년 4월 발표	• 멀티모달·긴맥락 대응 강화, 메타의 주력 모델
Gemma 3 2025년 3월 발표	• 효율 중심 오픈모델, 비용/운용 측면에서 추천
Llama 3.3 70B 2024년 12월 발표	• 기존 Llama 3.1 405B에 준하는 고성능을 훨씬 가벼운 70B 크기로 압축하여 효율성 극대화

3. AI 기술 발전 전망


대한민국 '국가대표 AI' 프로젝트: 소버린 AI 강국을 향한 여정
AI
프로젝트 핵심 목표 및 지원 체계
5,300억 원
규모의 집중 투자
2027년까지 GPU 인프라, 고통질 데이터 확보, 전문 인력 채용을 전목적으로 지원합니다.
GPT, Camini (글르벌 AI)
세계 최고 수준
글로벌 모델 대비
95% 성능 확보
GPT-4 및 Gemini 등 세계 최고 수준 모델의 95% 성능에 도달하는 것을 목표로 합니다.
GPU
GPU 인프라
생태계 구축
엔비디머큐부터 26만 장의 GPU를 우선 확보하여 소버린 AI 구축을 가속과합니다.
서바이벌 로드맵 및 경쟁 현황
2025
SK K1
LG AI인구원
초기 3개 팀 등탭
업스테이지
모드레이지
주요 핀소시입 현황 (1차 평가 통과 후)
컨소사업
모델
특버사항
SK텔래콤
A.X K1
국내 최대 크기 모델 공개 예정
LG AI연구원
K-EXAONE
LG 계열시 역한 금입 및 뫼토나실 구축
업스테이지
Solar Open 100B
글로릴 성능 김증 모델 기만 좡여
모티프테크 놀로지스
(추가 선정)
재차부협진을 흉해 경예임 힘류
6개월 주기의
단계별 탈락제
징기적인 중간 평가를 통해 경챙뵉뭄 검중하여, 머달 팀은 측시 제와되는 구조핍니다.
국민 평가단 도입
500영의 국민이 작평 모맵의 사용성뭄 평가하여 기술력 판섭도라 실용성을 모두 짐증합니다.
2027
AI
AI
최종 2개 팀 선발
2025년 5개 팀으로 시착하여, 최종적으로 국가를 대표할 2개 모탤만 납개 됩니다.
대한민국 국가대표 AI
소버린 AI 강국 달성
소버린 AI 강국 달성
NotebookLM

4. AI 인프라 구축의 주요 구성 요소

AI 인프라 구성

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

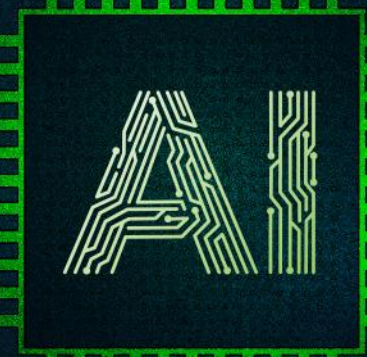
ESG를 위한 전력 효율화 솔루션

Arm 기반 친환경 서버
Liquid Cooling 냉각 방식

HS 효성인포메이션시스템 | HITACHI

2. 멀티에이전트 구축을 위한 AI플랫폼

1. 멀티에이전트 인프라 구성의 복잡성
2. HS효성 AI플랫폼
3. AI를 위한 데이터 인프라 아키텍처
4. AI 오케스트레이션 플랫폼 - 히타치 IQ 스튜디오
5. AI 도입 이슈에 대한 고민 해결



1. 멀티에이전트 인프라 구성의 복잡성

멀티에이전트 인프라 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/서버 가상화)

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?

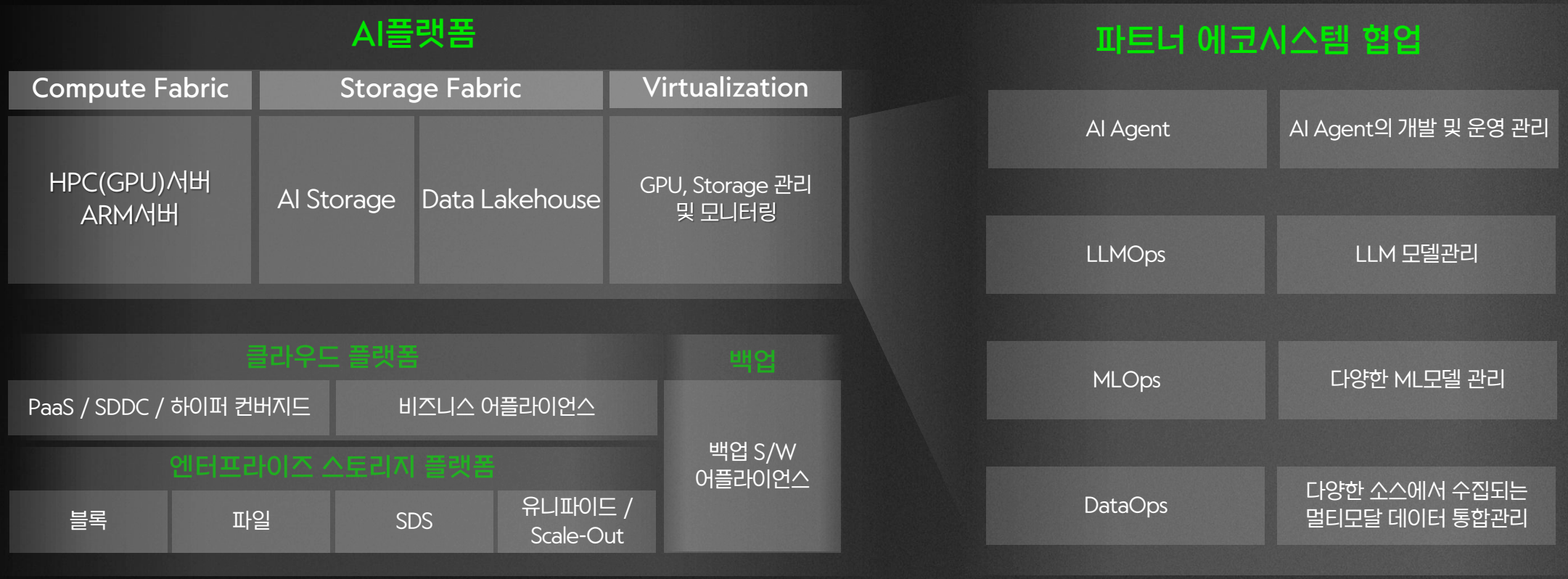
도입 후 활용은 ?

어떻게?

2. HS효성 AI플랫폼

확장성과 유연성을 갖춘 AI플랫폼

파트너 에코시스템 시너지 강화



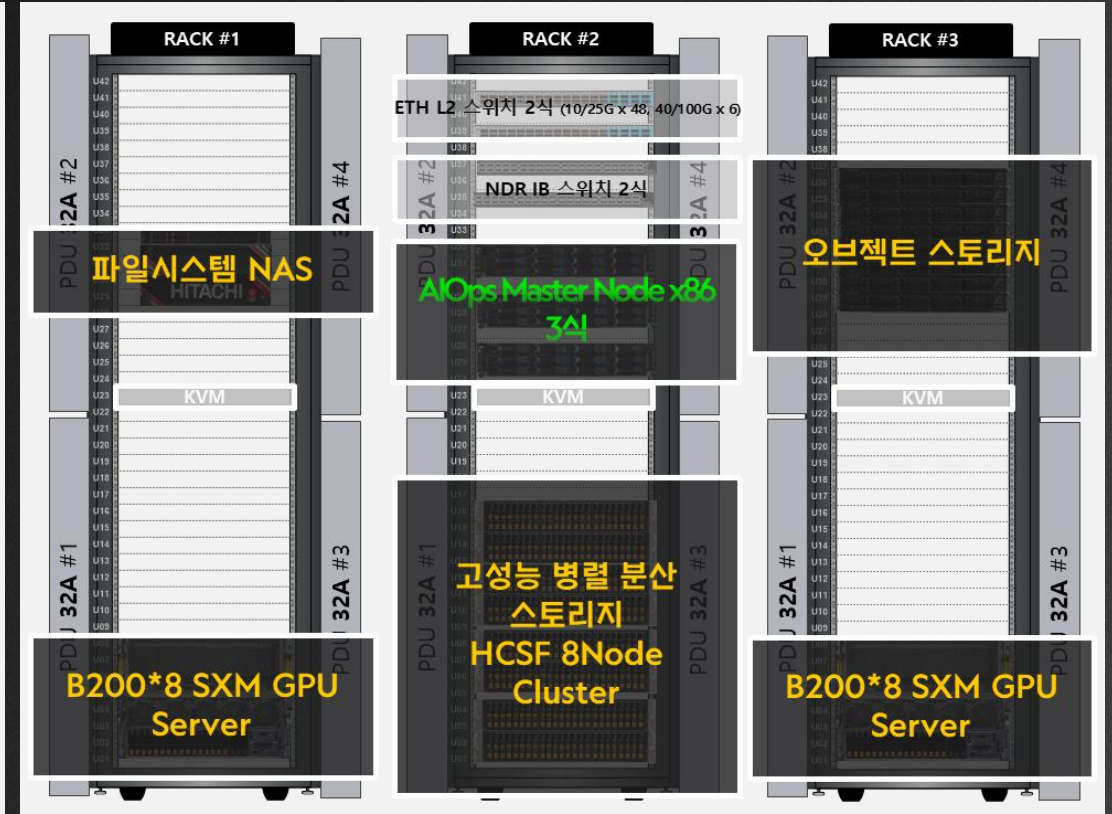
2. HS효성 AI플랫폼

확장성과 유연성을 갖춘 AI플랫폼 구성
GPU Infra + 고성능 Storage + 고성능 NW + AIOps SW

소규모 추론용 AI플랫폼

고성능 학습/추론용 AI플랫폼

초고성능 학습/추론용 AI플랫폼

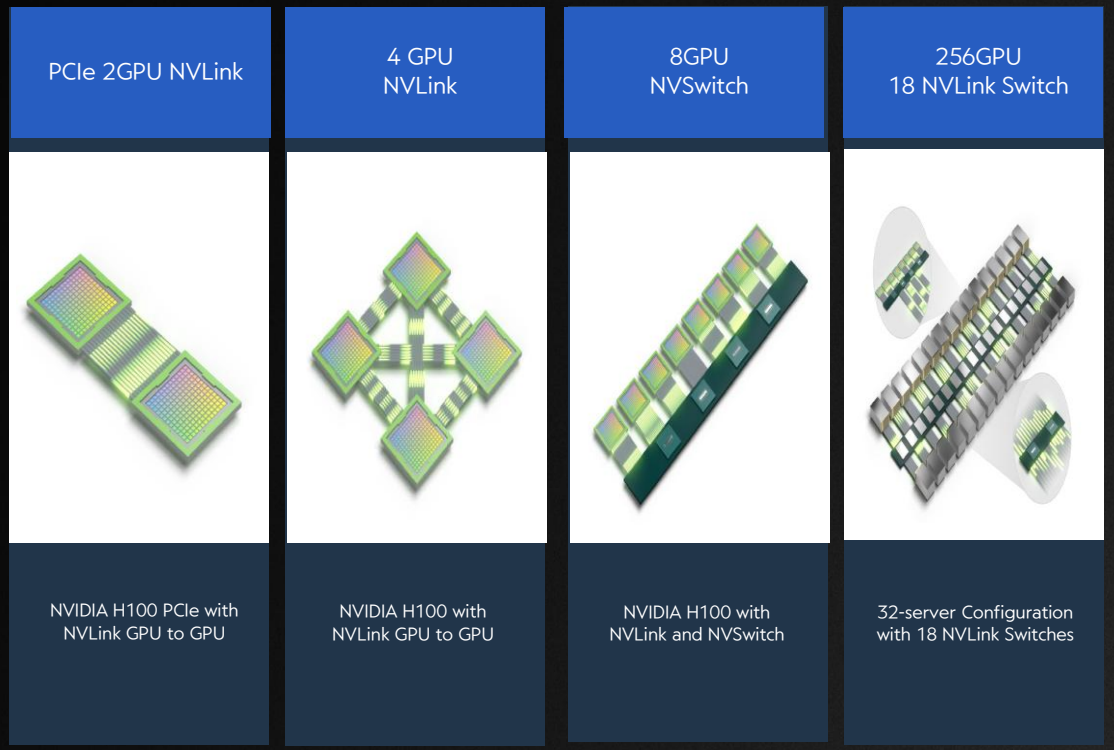


*상기 랙 실장도의 서버 이미지 및 수치, mount 크기(ex:2U → 1U)는 변경될 수 있습니다.

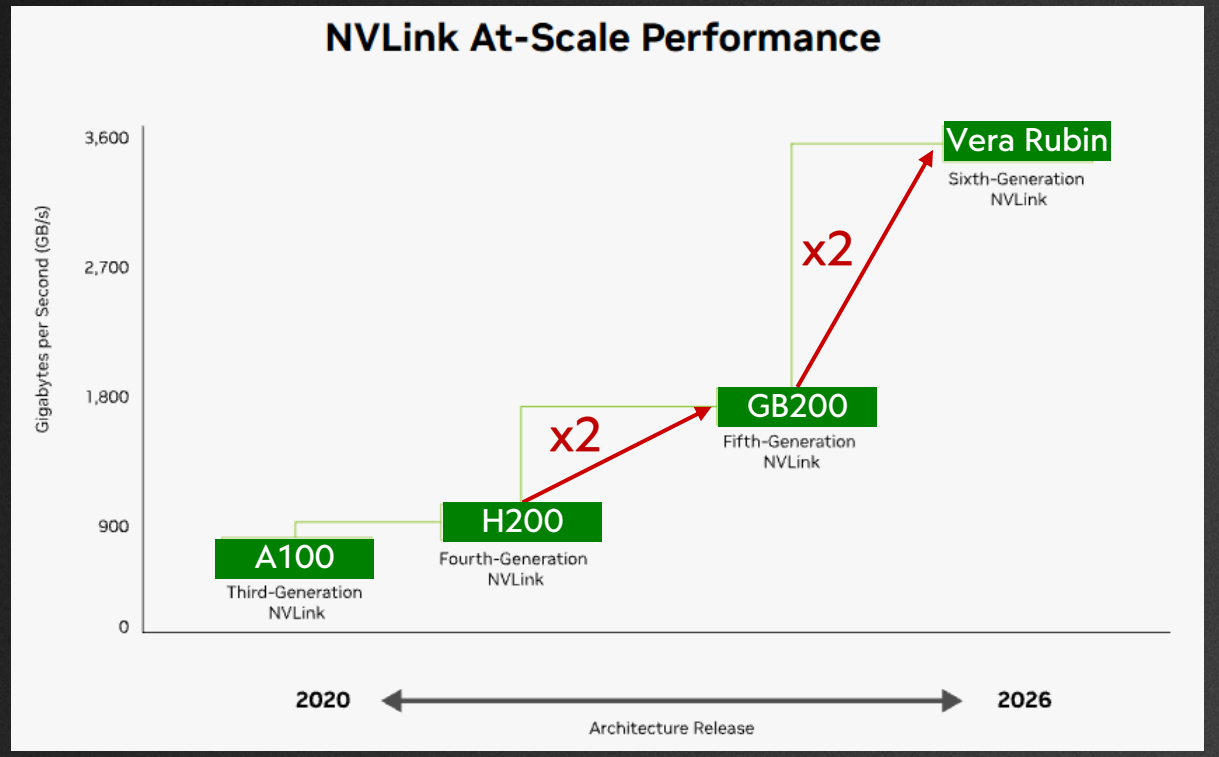
3. AI를 위한 데이터 인프라 아키텍처

- AI 인프라 도입 시 고려 사항_ **연산 성능 향상**

  고성능을 위한 HPC 컴퓨팅 환경 구축, 고성능 CPU 또는 GPU 팜 구성(서버, 네트워크, 스토리지)
  NVIDIA NVLink와 NVLink Switch(NVSwitch)는 여러 개의 GPU를 하나로 묶어 거대한 단일 GPU처럼 작동하게 만드는 기술



※ 출처: NVIDIA H100 Tensor Core GPU Architecture 자료



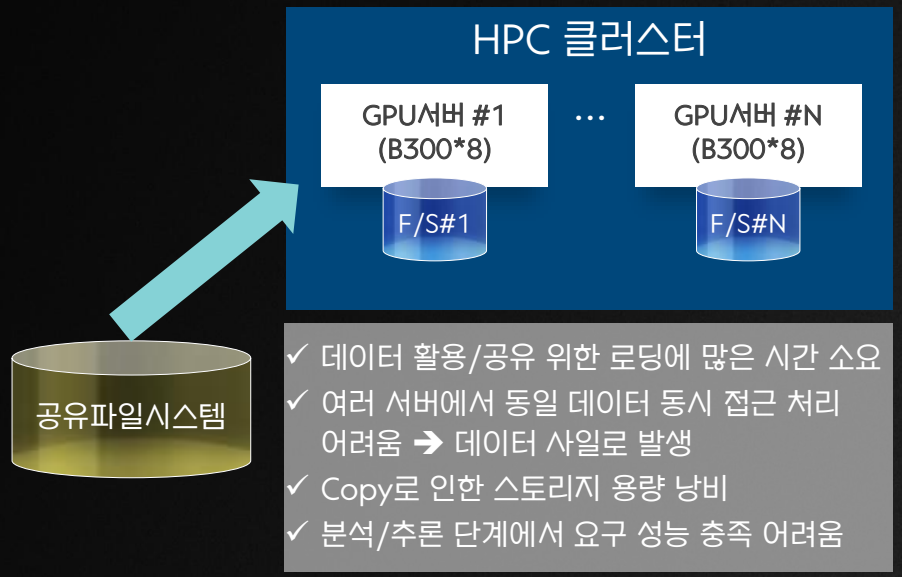
※ 출처: NVIDIA

3. AI를 위한 데이터 인프라 아키텍처

- AI 인프라 도입 시 고려 사항_ **스토리지 인프라 고속화**

  고성능 스토리지 도입(All Flash), 데이터 이동 시간 단축, GPU-스토리지간 성능 향상(GDS : GPUDirect Storage)
  데이터 수집에서 분석, 그리고 분석 후 데이터 공유 및 전달까지의 모든 경로를 고속화하고 이동을 최소화한 설계를 통해 전체 분석 시간 단축

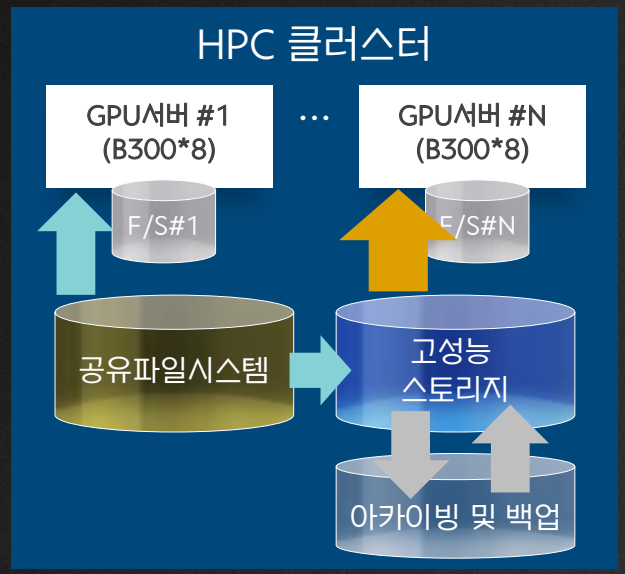
고성능 통합 파일 스토리지와 고속네트워크(인피니밴드 또는 100G 이상 이더넷 네트워크 구성) 구성을 통한 성능 향상



수집/저장 스토리지

단일 서버 고성능 디스크

GPU서버들 간 고속 네트워크 기술 (NVLink 등)



고성능 디스크, 고속 네트워크,
고성능 처리 기술 적용
(GDS/RDMA, DPDK, LocalIO 등)

통합 파일 시스템
고성능 처리 기술 적용
백업과 아카이빙 통합 구성

4. AI 오케스트레이션 플랫폼 - 히타치 IQ 스튜디오

- 기업용 AI 에이전트 구축·운영을 간소화하는 통합 플랫폼
- 노코드 에이전트 빌더와 온프레미스 보안 환경 제공

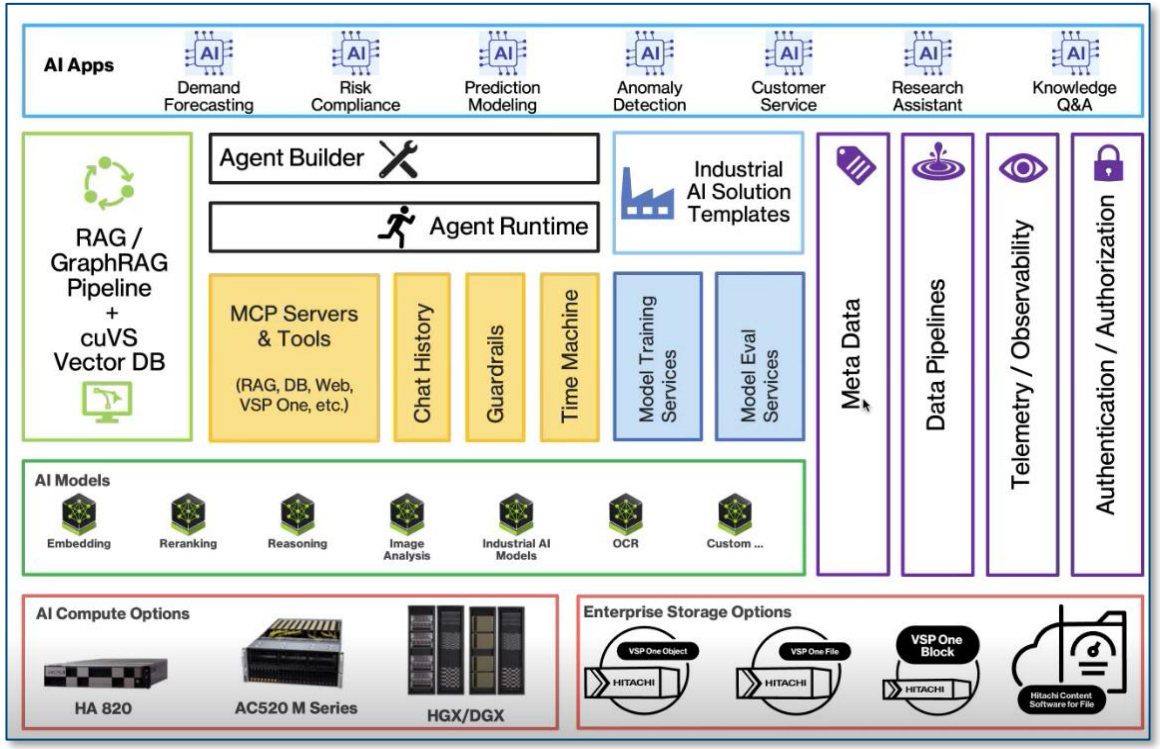
히타치 iQ 스튜디오

엔터프라이즈 AI 에이전트 구축/운영 간소화

'히타치 iQ 스튜디오'

- AI 에이전트 전 주기 통합 관리
- 엔비디아 AI 플랫폼 기반
- RAG/MCP 기반 AI-Ready 데이터제공
- AI 거버넌스 및 감사 추적 지원

히타치 iQ 스튜디오 아키텍처



5. AI 도입 이슈에 대한 고민 해결

1. AI 인프라 기술
- 통합 AI플랫폼 제공

- GPU 가상화, 고성능 스토리지, 네트워크, 컨테이너
- 슈퍼마이크로 GPU서버와 스토리지 조합으로 아키텍처 단순화

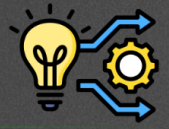
2. 비용효율적 구성
- 성능과 비용 효율 데이터 운영

- 고성능 데이터 처리 인프라 제공
- 초고성능 병렬 파일 스토리지 (Weka-HCSF)
- 고성능 파일 통합 스토리지(해머스페이스)
- 비용효율적 저장용 데이터레이크(오브젝트 스토리지)

3. 에코시스템 구축
- 다양한 솔루션 접목

- AI 적용을 위해 필요한 다양한 솔루션 접목
- 기존의 방식과 다른 접근 체계 가능
- AIOps, LLM 기반 챗봇 등 서비스 전문 파트너와 연계

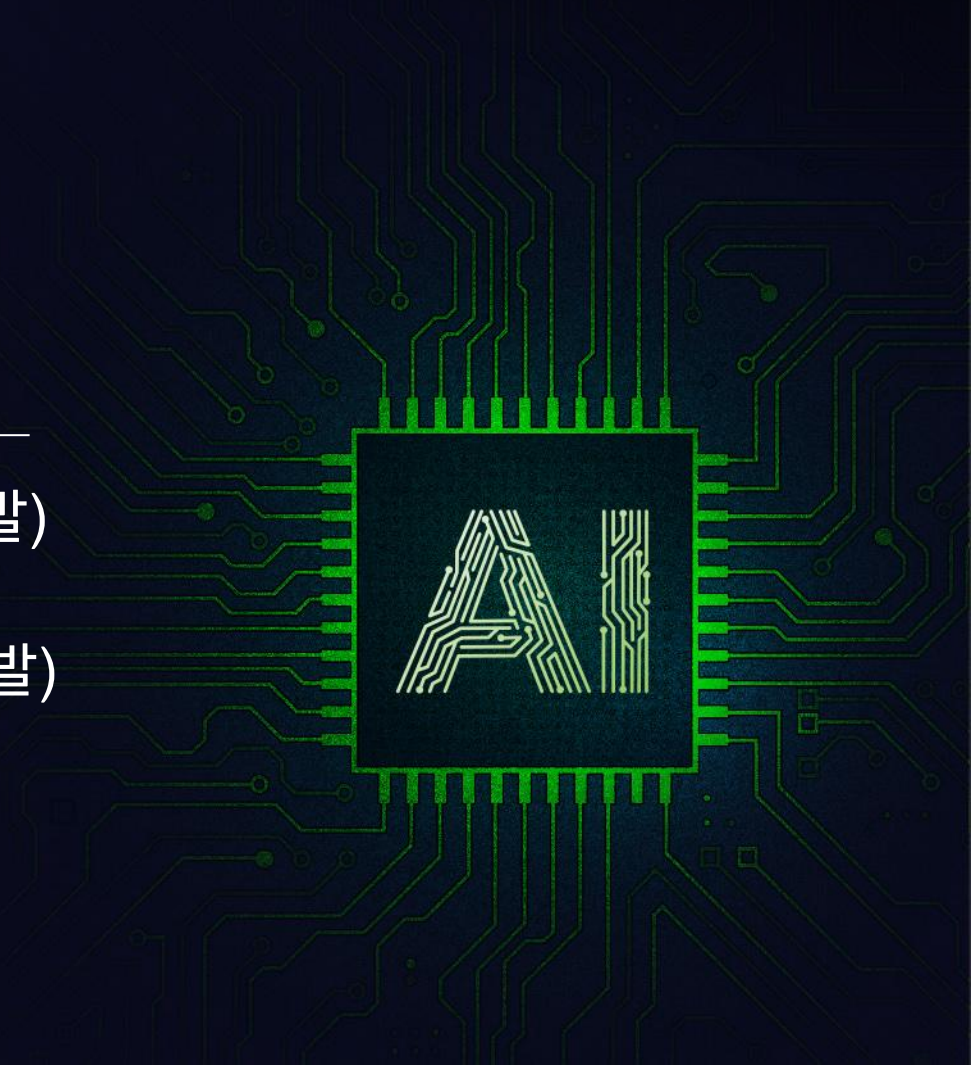
4. 운영 효율화
- 통합 제안 및 운영 지원



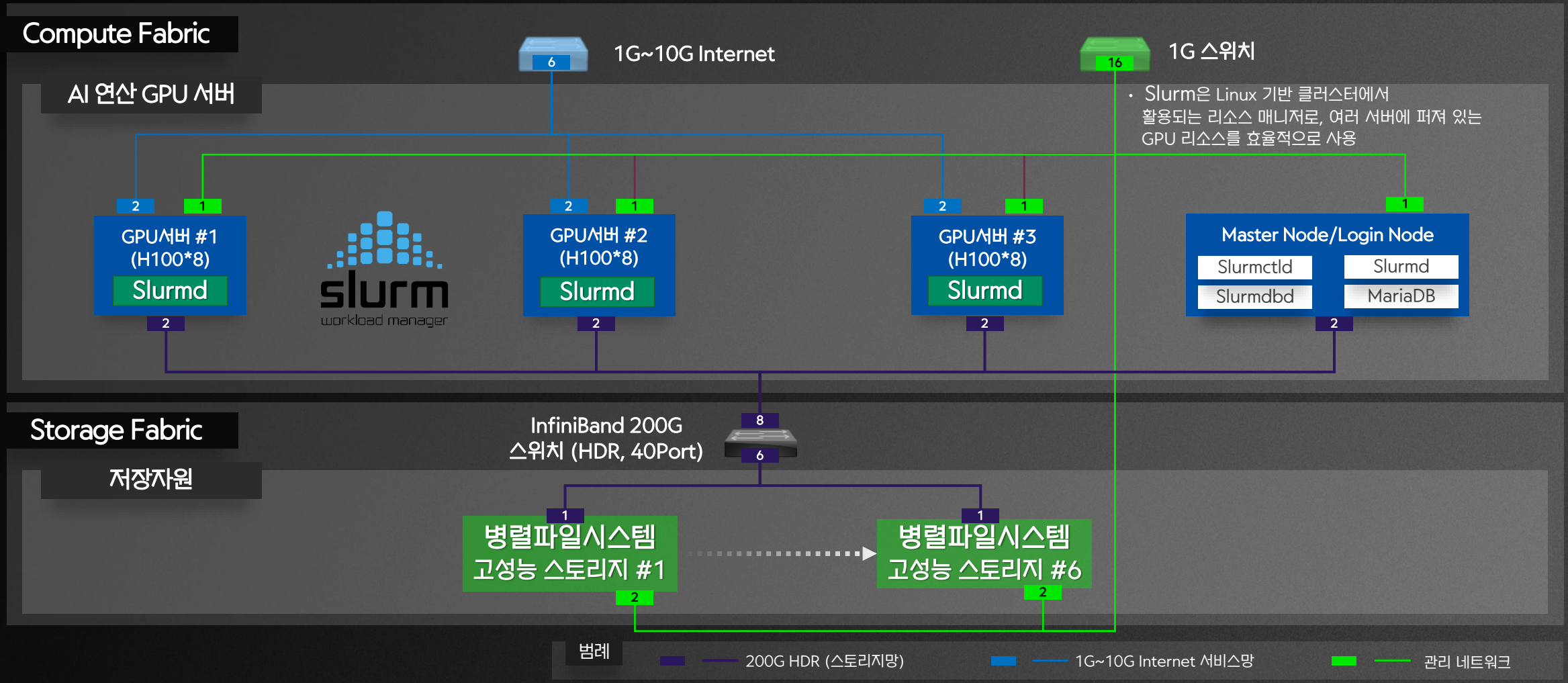
- AI 인프라에서 필수적인 연산자원과 (슈퍼마이크로 GPU서버) 네트워크, 저장자원 (SAN/NAS 및 HCSF, 해머스페이스, 오브젝트 스토리지 등)을 통합 구성
- 다양한 연계 솔루션을 통합 구축을 통해 운영 효율성 확보

3.
AI플랫폼 사례 및 구성

1. A사 사례-IT 대기업 AI플랫폼 인프라(자체 LLM 개발)
2. B사 사례-금융 데이터레이크 GPU AI 인프라 구축
3. C사 사례-대기업 DX GPU AI 인프라 구축 (연구 개발)
4. AI 인프라 도입 시 고려 사항



1. A사 사례-IT 대기업 AI플랫폼 인프라 (자체 LLM 개발)



2. B사 사례-금융권 데이터레이크 GPU AI 인프라 구축

사업 목적

- 외부기관 데이터/국내금융기관 데이터 통합으로 데이터 분석 및 AI 분석 /LLM 환경 확립

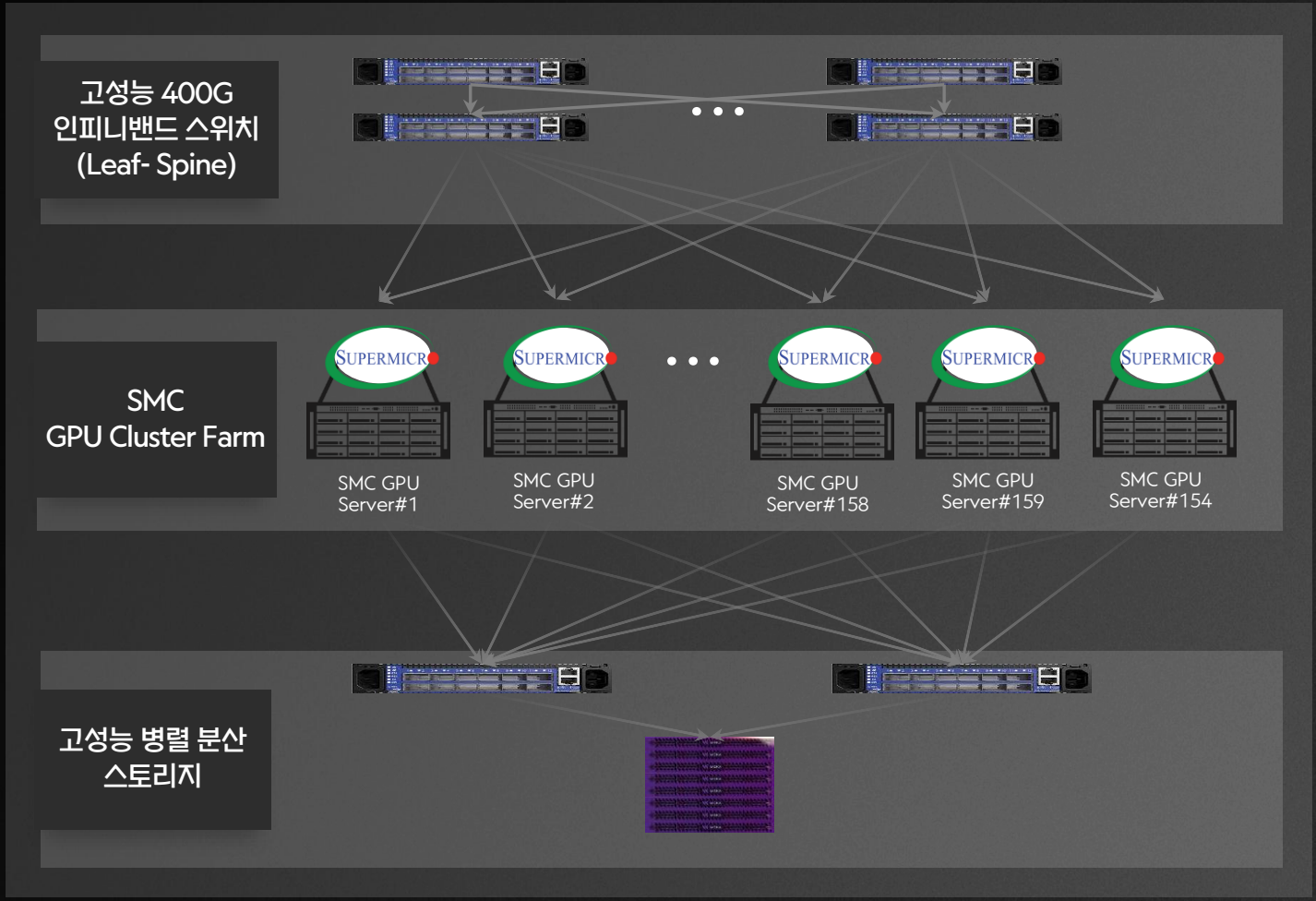
구축내용

- 정형/반정형/비정형 데이터 분석을 위한 데이터 레이크 구축
- GPU기반 AI 분석환경 및 LLM환경으로 Scale-out 아키텍처 수립

도입효과

- Row기반/Column 기반 분석이 모두 가능한 고성능 쿼리 분석 엔진 도입
- 통합 플랫폼 (CPU -> + GPU) 을 이용해 HW관리 포인트 최소화 및 확장 유연성 제공
- 고성능 단일 데이터 레이크 저장소 구축 운영

3. C사 사례-대기업 DX GPU AI 인프라 구축 (연구 개발)



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

구축 내용

- 제조 연구 개발 및 분석을 위한 AMD GPU Farm 구축
- ROCm 라이브러리 활용을 통한 분석 환경 및 병렬분산스토리지 연계

도입 효과

- One Vendor GPU (Nvidia) 종속성 탈피 및 cost saving을 위한 고성능 AMD GPU가 장착된 안정적인 고성능 SMC GPU Server 도입

4. AI 인프라 도입 시 고려 사항

01

AI플랫폼 구축 경험 확인

Compute Fabric,
Storage Fabric, AIOps Stack

02

국내외 실 사례를 통한 국내 기술력(인력) 여부 확인

장애 지원, 신규 AI 솔루션 연계 지원

03

다양한 에코 파트너 협업 체계 구축 확인

빠르게 변화하는 AI시대 대응

HS 효성인포메이션시스템 | HITACHI

감사합니다.