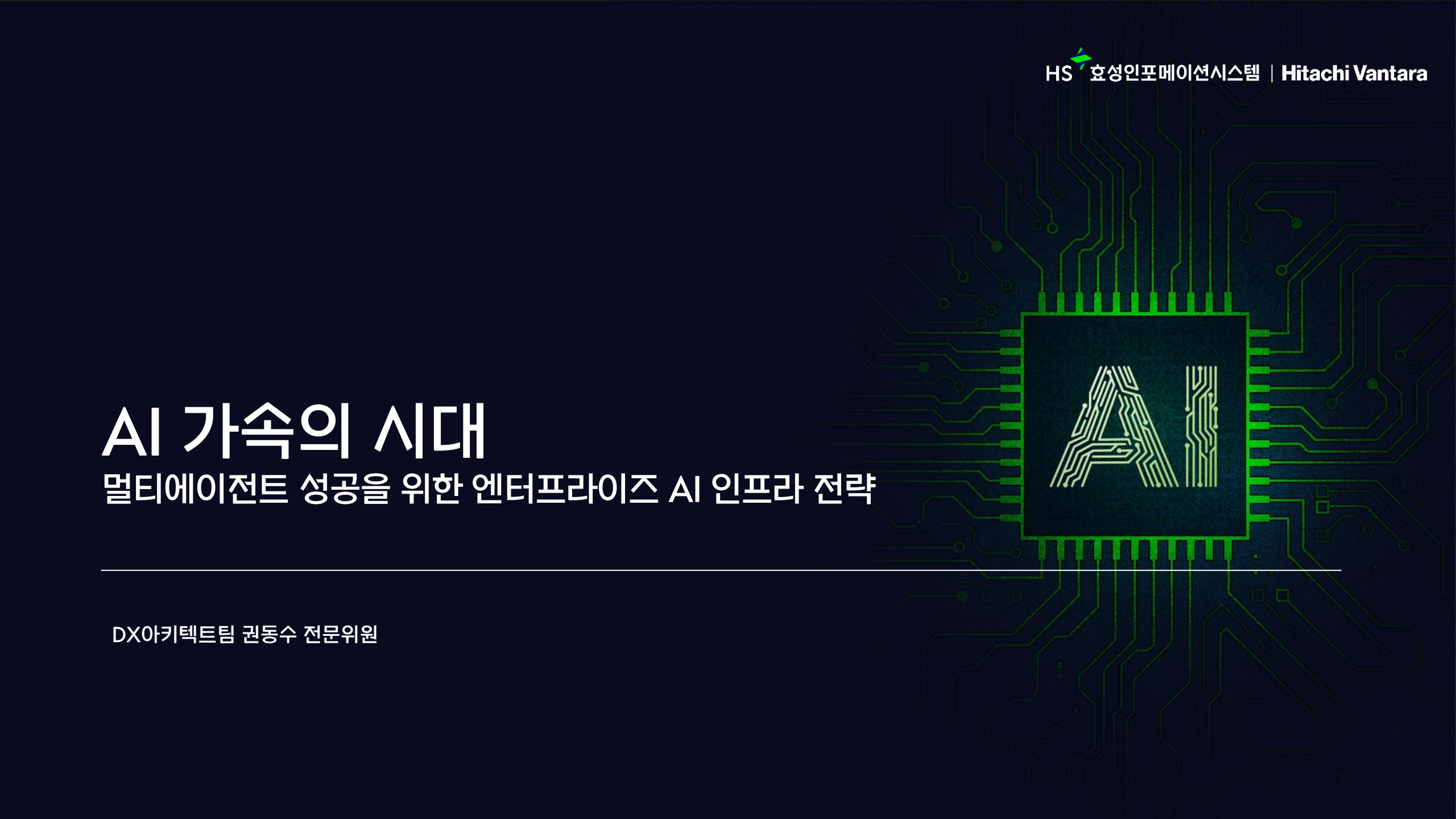
HS 효성인포메이션시스템 | Hitachi Vantara



AI 가속의 시대

멀티에이전트 성공을 위한 엔터프라이즈 AI 인프라 전략

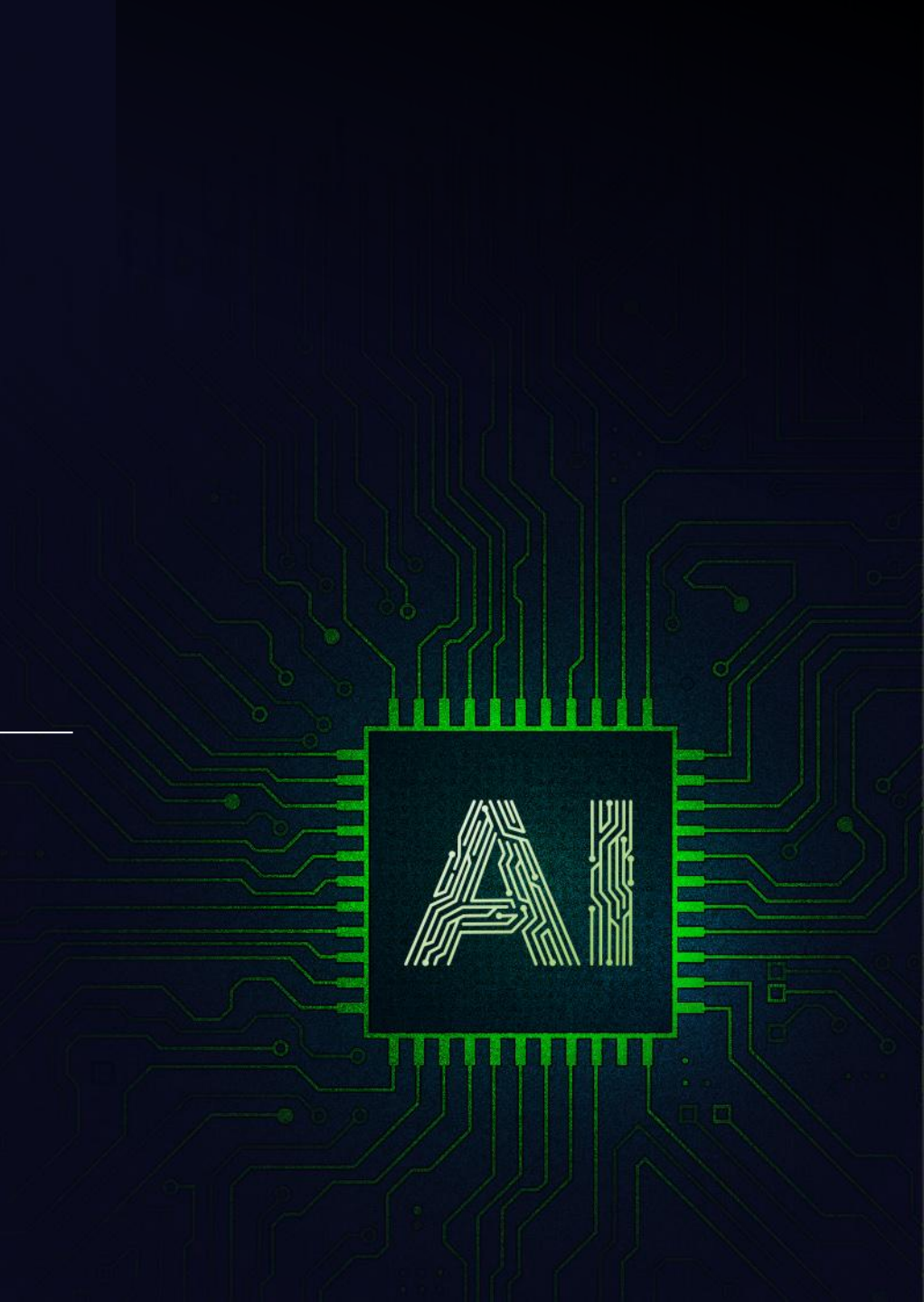
DX아키텍트팀 권동수 전문위원

Agenda

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

1.
AI 트렌드

1. 지속가능한 미래를 위한 AI
2. AI 기술 발전 전망



1. 지속가능한 미래를 위한 AI

멀티모달 AI

텍스트, 이미지, 음성 등
다양한 데이터를 동시에 처리

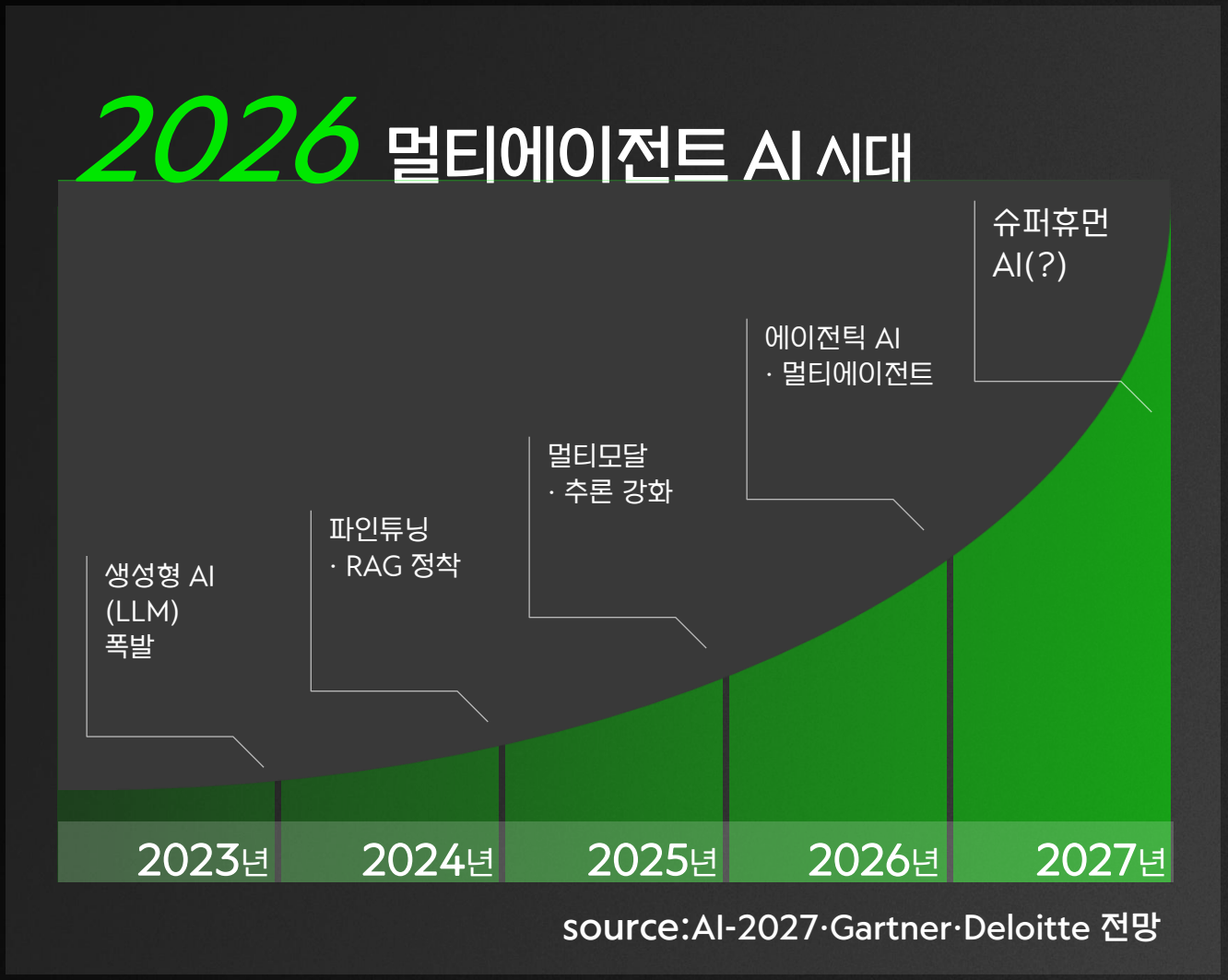
Agentic AI

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. AI 기술 발전 전망



모델명	특징 요약
Llama 4.5 (Enhanced) 2026년 1월 발표	• 10M 토큰의 초거대 컨텍스트와 네이티브 멀티모달 능력 및 에이전트 성능 최적화 (MXFP4 양자화 포맷 지원)
gpt-oss-120B 2025년 8월 발표	• 강력한 추론 성능을 제공하는 전문가 혼합(MoE) 구조의 AI 모델(MXFP4 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
Llamma 3 70B 2024년 4월 발표	• 700억 개의 매개변수를 기반으로 추론, 코딩, 창의적 글쓰기에서 강력한 성능을 발휘
Mixtral 8×22B 2024년 4월	• 오픈소스 전문가 혼합(MoE) 모델 중 성능·효율 균형 우수상용

2. AI 기술 발전 전망

대한민국 '국가대표 AI' 프로젝트: 소버린 AI 강국을 향한 여정

프로젝트 핵심 목표 및 지원 체계

5,300억 원 규모의 집중 투자
2027년까지 GPU 인프라, 고품질 데이터 확보, 전문 인력 채용을 전폭적으로 지원합니다.

글로벌 모델 대비 95% 성능 확보
GPT-4 및 Gemini 등 세계 최고 수준 모델의 95% 성능에 도달하는 것을 목표로 합니다.

GPU 인프라 생태계 구축
엔비디어류부터 26반 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 컨소시입 현황 (1차 평가 통과 후)

컨소사업	모델	특버사항
SK텔레콤	A.X K1	국내 최대 크기 모델 공개 예정
LG AI연구원	K-EXAONE	LG 계열시 역한 곰입 및 뫼토나실 구축
업스테이지	Solar Open 100B	글로릳 성능 김증 모델 기만 쟝여
모티프테크 놀로지스	(추가 선정)	제차부컬진을 흉해 경예임 힐류

6개월 주기의 단계별 탈락제
징기적인 중긴 평가를 통해 경챙복을 검중하여, 머달 팀은 측시 제화되는 구조입니다.

국민 평가단 도입
500명의 국민이 작평 모멤의 사용성물 평가하여 기술력 란섭도라 실용성을 모두 짐증합니다.

2027

최종 2개 팀 선발
2025년 5개 팀으로 시칵하여, 최종적으로 국가를 대표할 2개 모탤만 납개 됩니다.

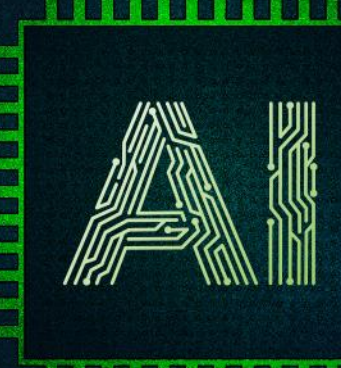
대한민국 국가대표 AI 소버린 AI 강국 달성

소버린 AI 강국 달성

NotebookLM

2.
멀티에이전트 구축을 위한 HS효성 AI플랫폼

1. AI 인프라 구성의 복잡성
2. HS효성 AI플랫폼
3. AI 오케스트레이션 플랫폼 - 히타치 IQ 스튜디오
4. 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?

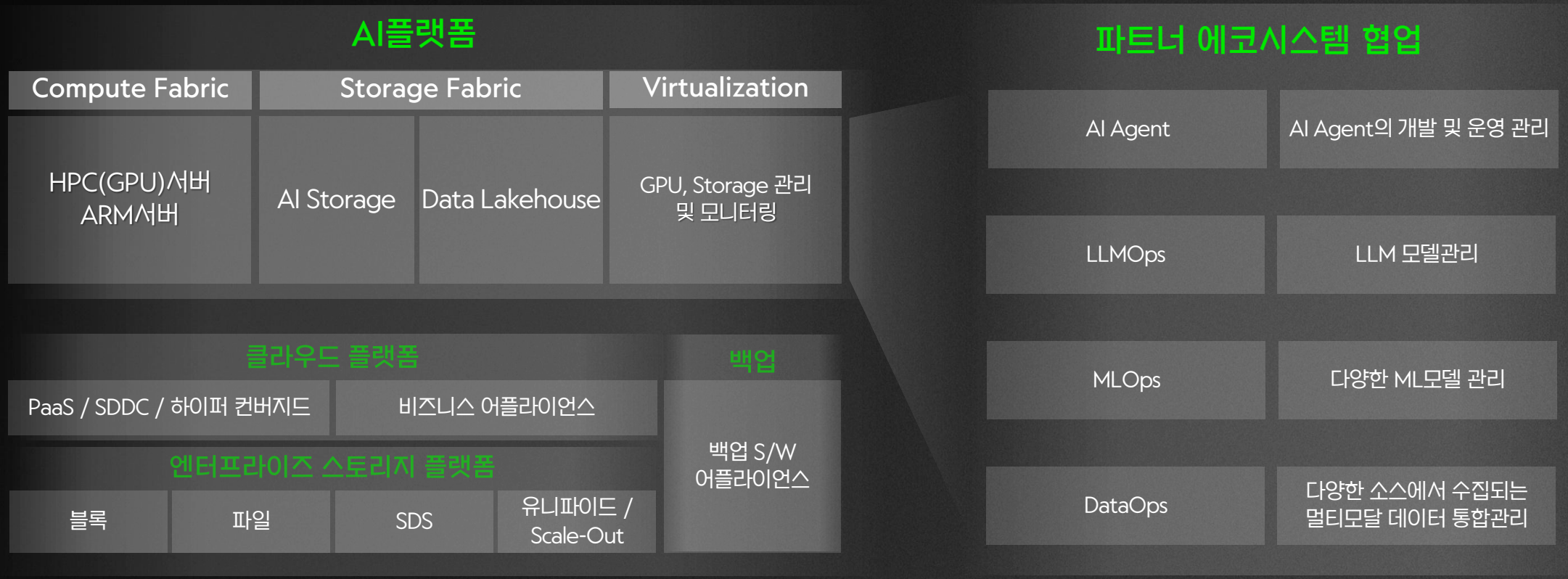
도입 후 활용은 ?

어떻게?

2. HS효성 AI플랫폼

확장성과 유연성을 갖춘 AI플랫폼

파트너 에코시스템 시너지 강화



3. AI 오케스트레이션 플랫폼 - 히타치 iQ 스튜디오

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

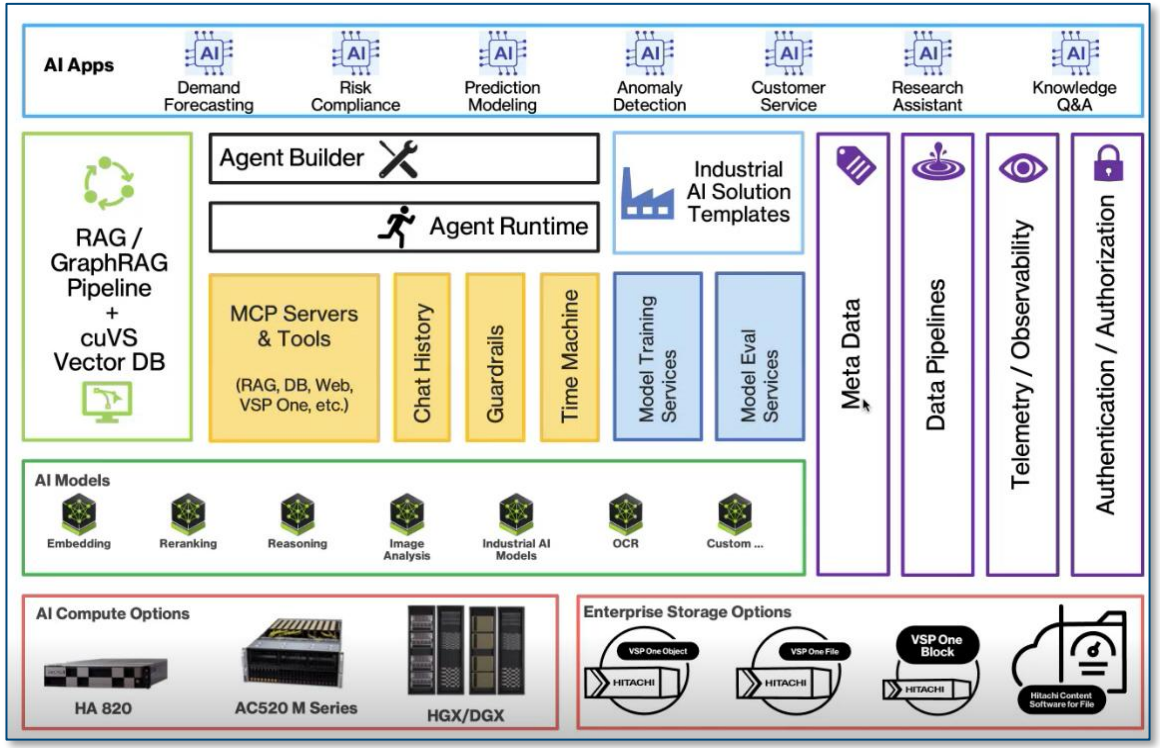
히타치 iQ 스튜디오

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



4. 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

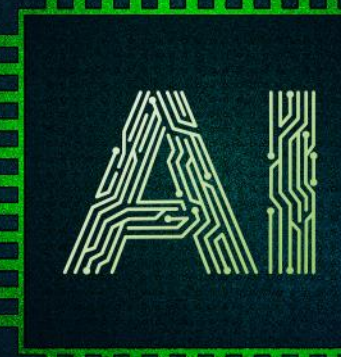
- AI 적용을 위해 필요한 다양한 솔루션 접목
- 기존의 방식과 다른 접근 체계 가능
- AI Ops, LLM 기반 챗봇 등 서비스 전문 파트너와 연계

4. 운영 효율화
- 통합 제안 및 운영 지원

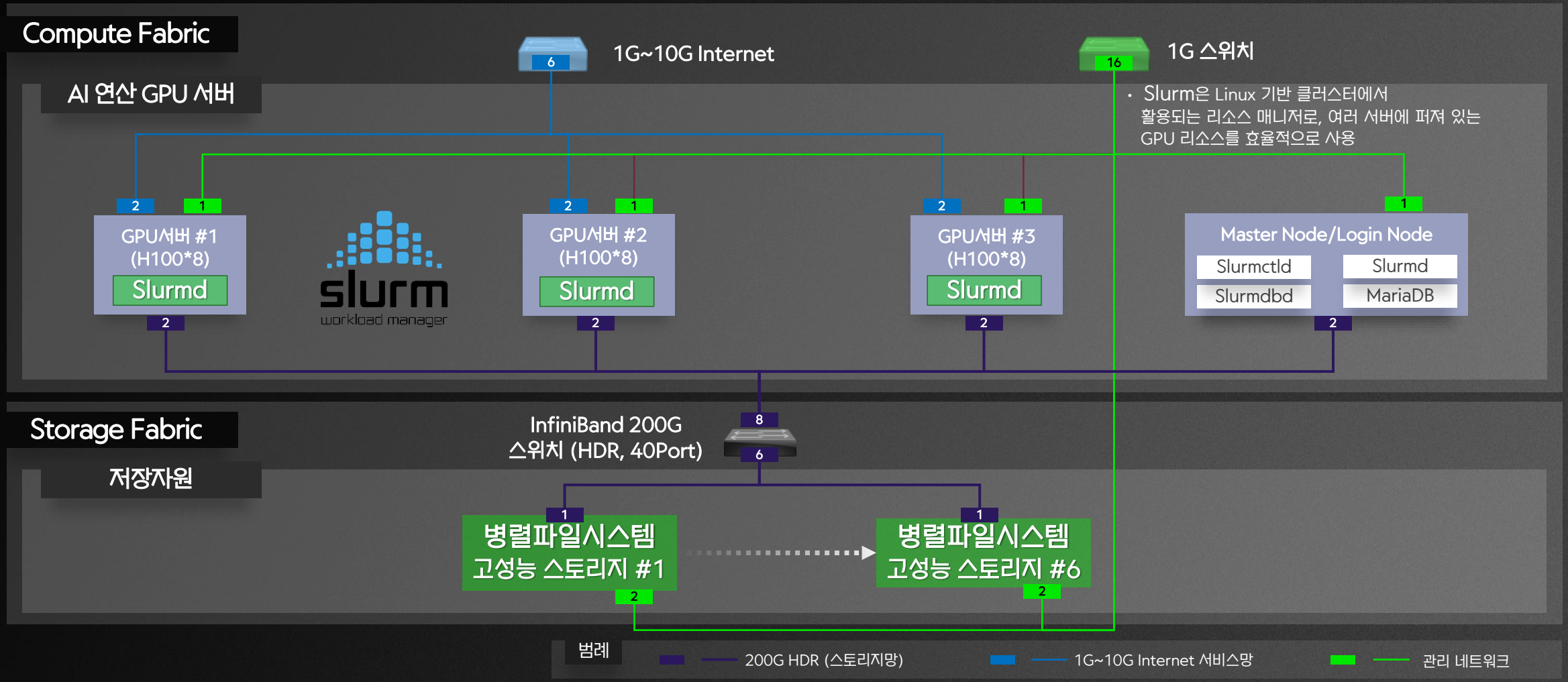
- AI 인프라에서 필수적인 연산자원과 (슈퍼마이크로 GPU서버) 네트워크, 저장자원 (SAN/NAS 및 HCSF, 해머스페이스, 오브젝트 스토리지 등)을 통합 구성
- 다양한 연계 솔루션을 통합 구축을 통해 운영 효율성 확보

3.
AI플랫폼 사례 및 구성

1. A사 사례-IT 대기업 AI플랫폼 인프라(자체 LLM 개발)
2. B사 사례-대학병원 AI 분석 플랫폼 GPU FARM 구축
3. C사 사례-대기업 DX GPU AI 인프라 구축 (연구 개발)
4. HS효성 AI플랫폼 구성
5. AI 인프라 도입 시 고려 사항

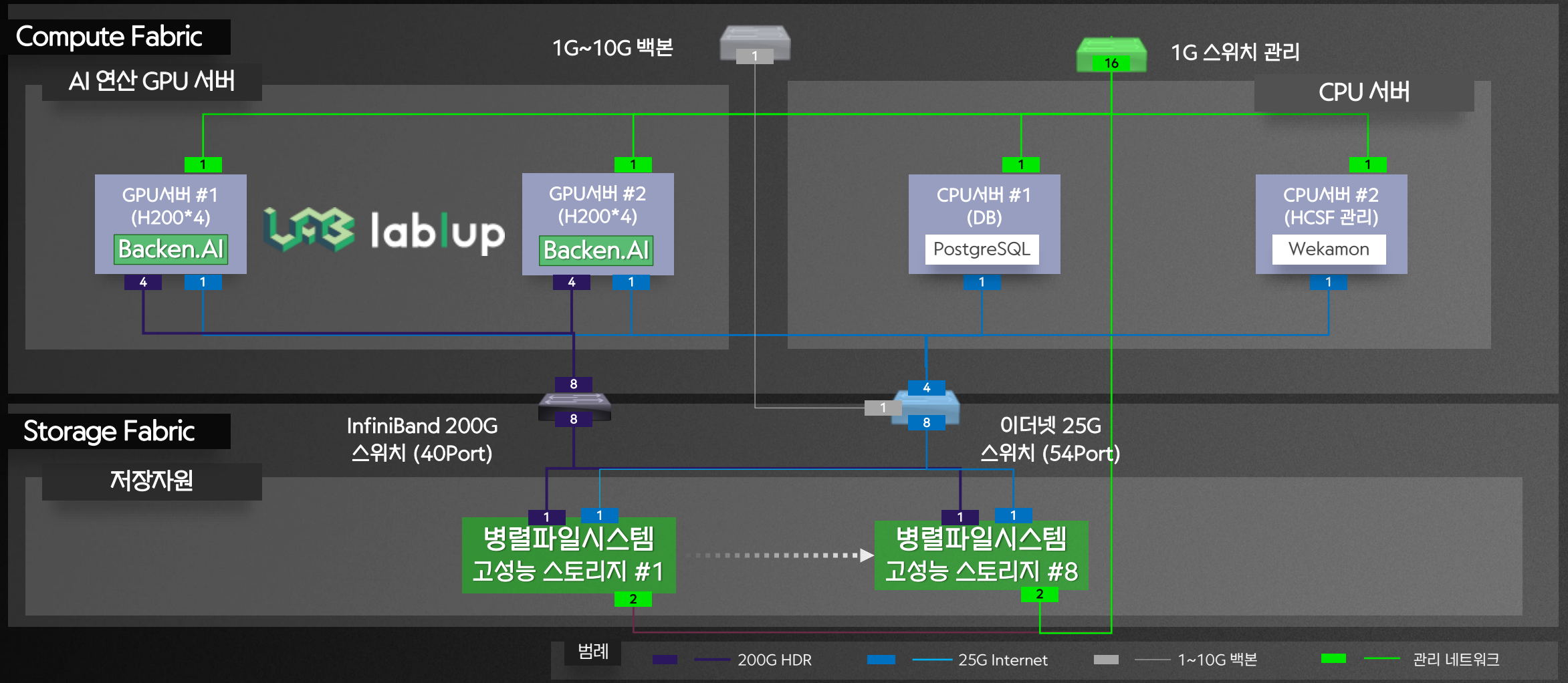


1. A사 사례-IT 대기업 AI플랫폼 인프라(자체 LLM 개발)

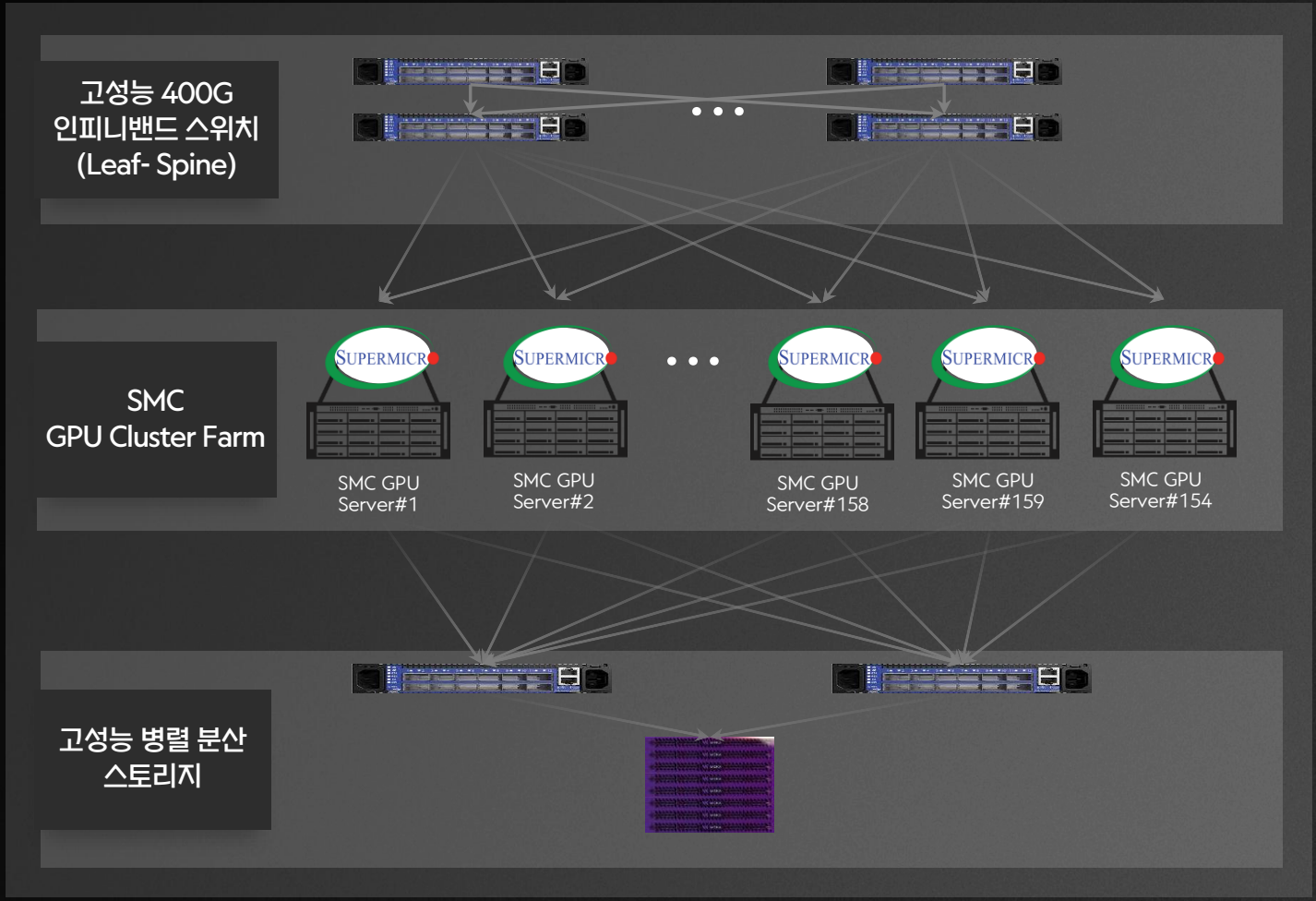


2. B사 사례-대학병원 AI 분석 플랫폼 GPU FARM 구축

HPC 클러스터 (HW & SW) & HCSF(분석특화 저장소) & GPU 가상화 분할 및 클러스터 관리 솔루션 구축 사례



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

SuperMicro + HIS 선정 이유

- SuperMicro Server의 모듈식 설계에 따른 유연성과 확장성, 서버 성능을 보장하는 안정성
- HIS의 전문 기술지원 인력을 통한 최고의 직접 기술지원 서비스와 HPC/고성능 스토리지/AI 구축 레퍼런스

AMD GPU 선정 이유

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

4. HS효성 AI플랫폼 구성

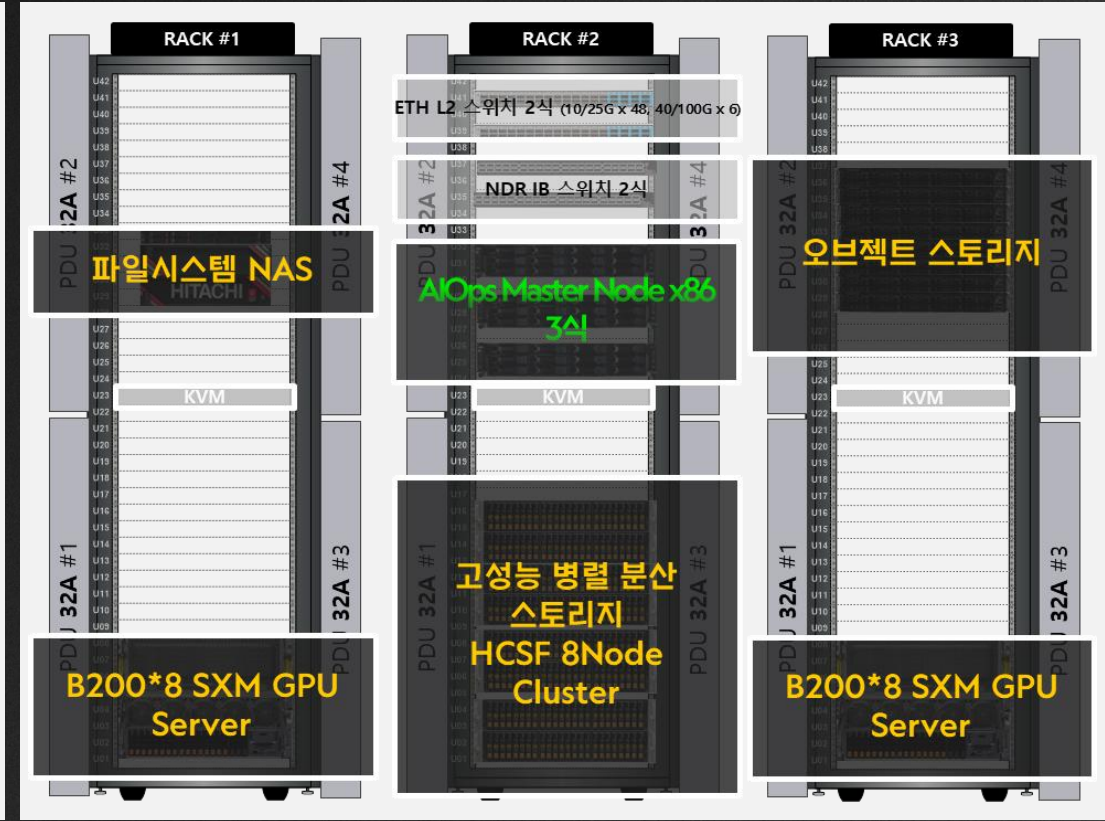
확장성과 유연성을 갖춘 AI플랫폼 구성

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.

5. AI 인프라 도입 시 고려 사항

01

다양한 에코 파트너 협업 체계 구축 확인

빠르게 변화하는 AI시대 대응

02

AI플랫폼 구축 경험 확인

Compute Fabric,
Storage Fabric, AIOps Stack

03

국내외 실 사례를 통한 국내 기술력(인력) 여부 확인

장애 지원, 신규 AI 솔루션 연계 지원

감사합니다.