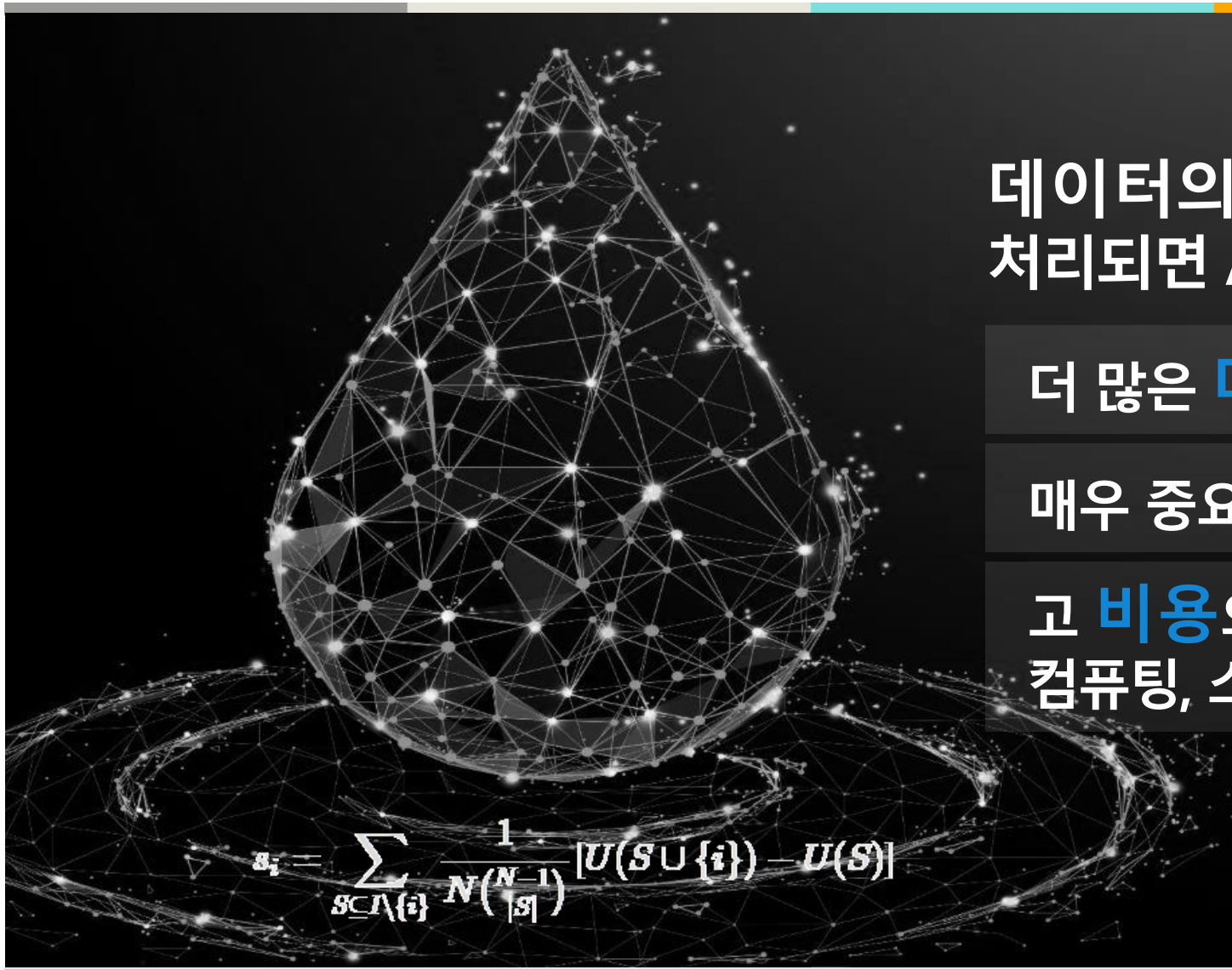
전자신문인터넷 인텔리전트 하이퍼오토메이션 2022

AI 업무 환경 확대, 시작부터 활용까지 한번에 이해하기

- 김 형 섭 컨설턴트

효성인포메이션시스템

AI 업무환경 확대를 위한 고민



데이터의 가치!!
처리되면 AI, ML 분석의 연료가 됩니다.

더 많은 **데이터** = 더 나은 결과,

매우 중요해진 **연산 속도**,

고 **비용**의 복잡한
컴퓨팅, 스토리지, 네트워크,

최적화 방법은 무엇입니까?

AI 업무환경의 확대

자동차 & 로봇

- 자율 주행
- 위험 및 규정 준수
- 로봇공학

사물인식(고급), 강화학습

생산 및 제조

- 수요 예측
- 에너지 소비 최적화
- 스케줄링 최적화
- 인재이탈 예측

강화학습, 추천

유지 관리 및 품질

- 예지 정비
- 예측 품질
- 시각적 분석

사물인식, 추천,
이미지 분할

유통 및 물류

- 화물 최적화
- 유통경로 최적화
- 선박운송 최적화

추천, 이미지 분류

헬스케어

- 의료영상 진단 분석
- 맞춤 진료처방 챗봇
- 치료 연구개발

이미지 분할, 자동음성인식,
자연어 처리, 사물인식(고급)

금융 & 서비스

- 보험상품 및 고객 분석
- 위조 및 사기 탐지
- 이상거래 탐지
- 고객상담 챗봇

강화학습, 자연어 처리,
자동음성인식

영업 및 마케팅

- 고객 분석
- 제품 가격 최적화
- 상품 추천
- 고객상담 챗봇

자연어 처리, 추천,
이미지 분류,자동음성인식

공공 분야

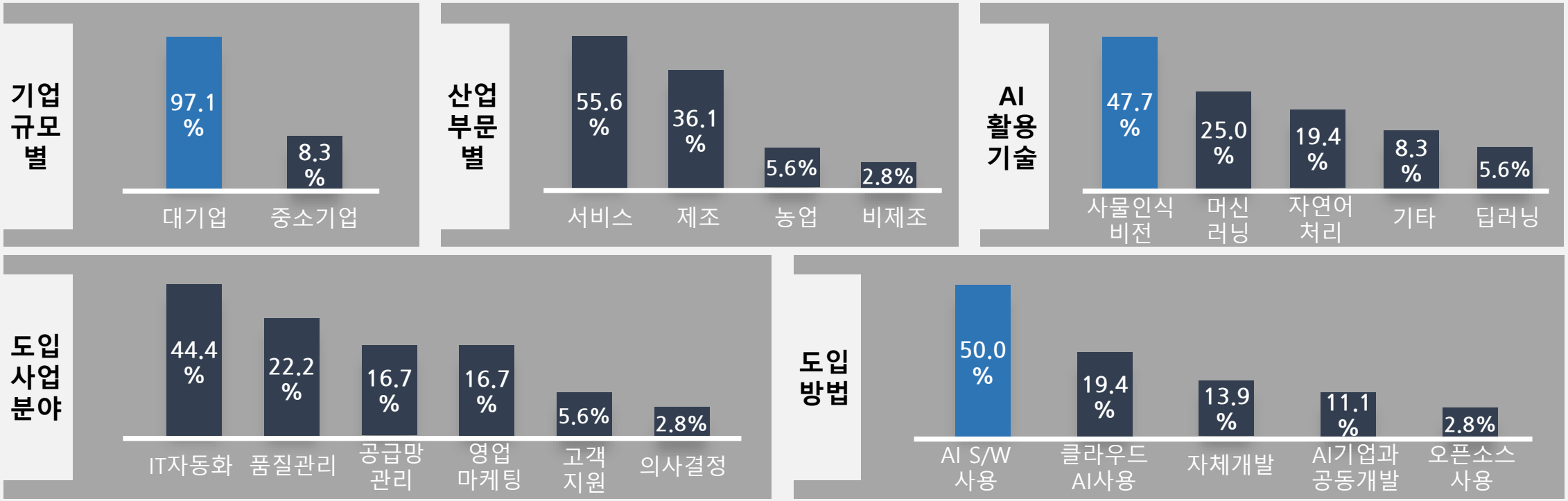
- 지능형 cctv
- 지능형 범죄 예방
- 공공 서비스 챗봇
- 연구 프로젝트

사물인식(일반), 추천
자동음성인식, 강화학습

국내 AI 도입현황 분석 (KDI 경제정보센터 여론분석팀 자료)

한정된 분야 AI 일부 적용에서, 전통적 업무 영역까지 AI 적극 도입 증가

대기업 중심으로 AI 기업용 솔루션 도입을 통해 AI의 비즈니스 적용 시도 확대

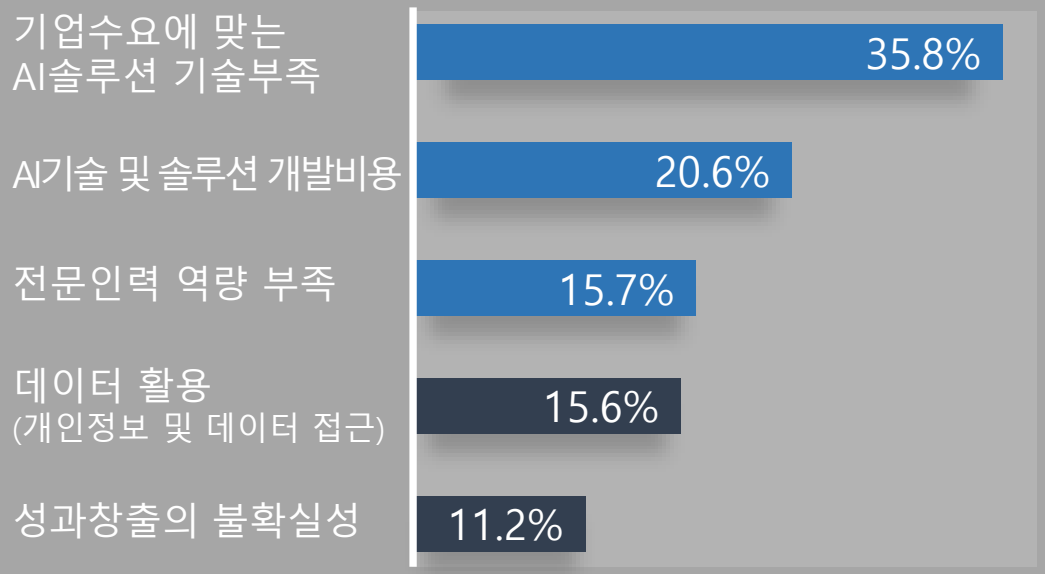


국내 AI 도입 이슈 (KDI 경제정보센터 여론분석팀 자료)

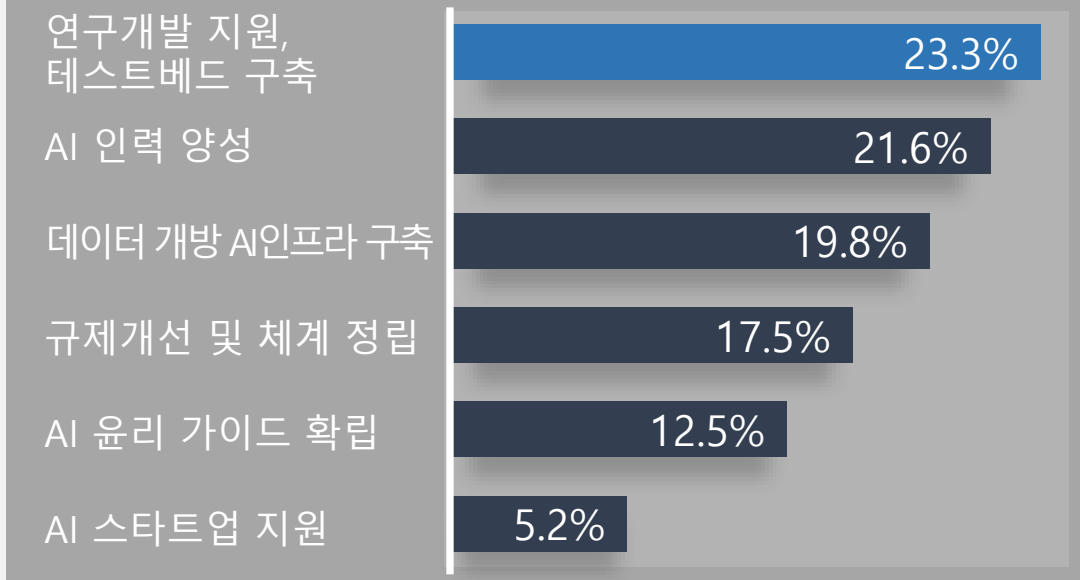
AI 기술, 비용, 역량에 대한 이슈로 AI의 비즈니스 적용의 어려움을 가지고 있음

성과 창출의 불확실성을 제거를 위한, AI 테스트 베드 구축 요구사항 증가

AI 도입의 걸림돌



AI 활성화를 위해 필요한 정책



국내기업 AI 도입 주요이슈

이슈 1. AI솔루션 기술 부족

. AI플랫폼은 복잡한 인프라 및 솔루션 조합으로 구성
(모델링 알고리즘, 클라우드, 컨테이너, GPU/서버 가상화)

이슈 2. 도입 비용 부족

. 고 사양의 H/W 인프라에 더해 AI 솔루션에 대한 비용 부담
(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?

AI도입 이슈에 대한 해답, 효성의 AI플랫폼

1. 기술
– 복잡한 AI 기술 솔루션의 통합 플랫폼 제공

- GPU서버, 고성능 병렬파일 스토리지, 네트워크, 컨테이너, GPU 가상화 기술의 통합 플랫폼 제공
- 복잡한 솔루션 조합이 아닌 슈퍼마이크로 GPU서버와 Backend.AI 솔루션 조합으로 아키텍쳐 단순화

2. 비용
– 성능과 비용 효율 데이터 운영

- 초고성능 병렬파일 스토리지 (Weka IO-HCSF) 를 통한 고성능 데이터 운영 인프라 도입
- Cold 데이터 오브젝트 스토리지 티어링 기반 저장효율

3. 역량
– 손쉬운 시작 및 활용

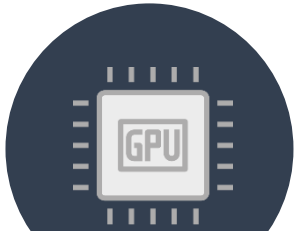
- Tensorflow 및 Pytorch 등 사전정의 개발환경 제공으로 연구환경 즉시 생성
- 직관적 사용자 UI로 효율적 AI모델 개발 및 컨테이너 운영

4. 구축
– 효성 단일 벤더 통합 오퍼링

- 연산자원과 (NVIDIA DGX, Supermicro HGX) 저장자원의 (HCSF) H/W 인프라와 함께 컨테이너 기반 GPU가상화 운영 솔루션 (Lablup Backend.AI) 통합제안으로 경쟁사 차별화

AI 업무 프로세스

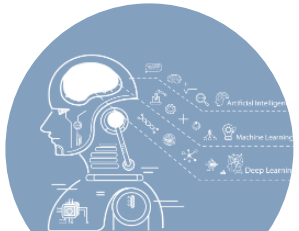
AI 플랫폼의 3가지 요소



연산 자원

학습 및 추론 연산자원

GPU



AI 모델

AI 알고리즘 개발환경

모델개발 환경

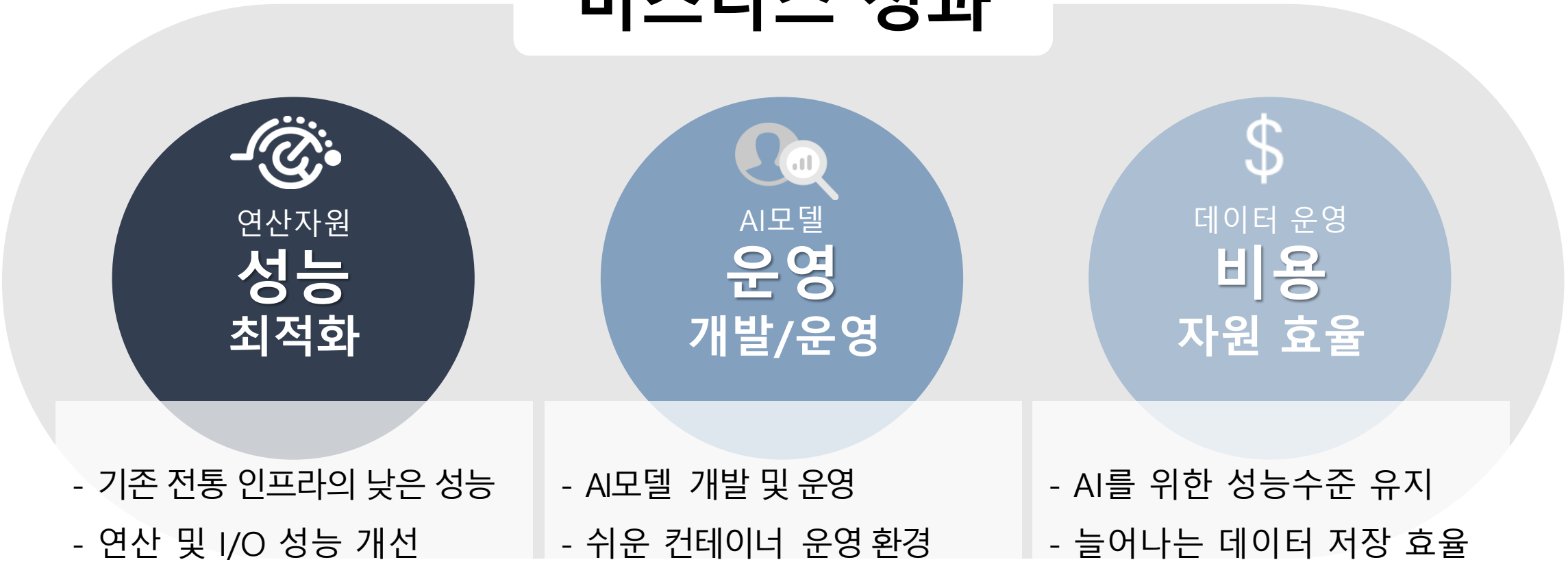
데이터

데이터 저장/분류/가공

저장 자원

AI 플랫폼의 필수 검토항목

비즈니스 성과



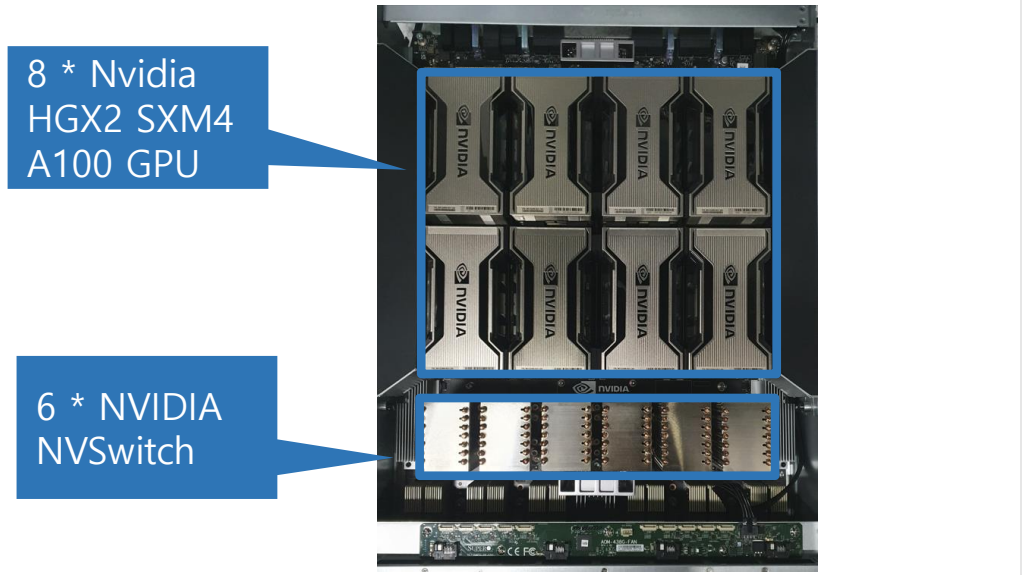
- 기존 전통 인프라의 낮은 성능
- 연산 및 I/O 성능 개선

- AI모델 개발 및 운영
- 쉬운 컨테이너 운영 환경

- AI를 위한 성능수준 유지
- 늘어나는 데이터 저장 효율

1. 연산자원 성능

NVIDIA DGX / HGX



- 4세대 NVIDIA NVLink는 900GB/s GPU 대역폭으로 PCIe Gen4 대비 14배 고성능
- NVIDIA A100 Tensor 코어 GPU의 고속 상호 연결 구현

NVLink와 PCIe 비교

항목	서브 링크 속도	서브 링크 수	전체 속도	GPU 아키텍쳐
PCIe 3.x	16GB/s	1	**16GB/s**	Pascal , Volta , Turing
PCIe 4.0	32GB/s	1	**64GB/s**	Volta, Ampere
NVLink 1.0	20GB/s	4	**160GB/s**	Pascal
NVLink 2.0	25GB/s	6	**300GB/s**	Volta
NVLink 3.0	25GB/s	12	**600GB/s**	Ampere
NVLink 4.0	25GB/s	18	**900GB/s**	Hopper

- GPU-GPU, CPU-GPU간 전송 기술로 GPU메모리 직접 통신
- NVLink 1세대에서 NVLink 4세대로 NVLink 방식 발전

1. 연산자원 성능

효성인포메이션시스템은 **슈퍼마이크로 공식 총판사**로서 GPU 및 x86 서버를 공급 합니다.

GPU 서버 with NVLink

AS -4124GO-NART+

8 x A100 GPU (4U)

2 x AMD CPU
Max 8TB Memory
6 x 2.5″ Drive bays
8 x PCIe 4.0 x16 LP
AIOM support
4 x 3000W PSU

AS -2124GQ-NART+

4 x A100 GPU (2U)

2 x AMD CPU
Max 8TB Memory
4 x 2.5″ Drive bays
4 x PCIe 4.0 x16 LP
1 x PCIe 4.0 x8 LP
Dual 1GbE NIC
2 x 3000W PSU

PCIe GPU 서버

AS -4124GS-TNR

8(10) x GPU (4U)

2 x AMD CPU
Max 8TB Memory
24 x 2.5″ Drive bays
Dual 1GbE NIC
AIOM support
4 x 2000W PSU

SYS-220GP-TNR

6 x GPU (2U)

2 x Intel CPU
Max 4TB Memory
10 x 2.5″ Drive bays
2 x PCIe 4.0 x8 LP
AIOM support
2 x 2600W PSU

SYS-740GP-TNRT

4 x GPU (4U)

2 x Intel CPU
Max 4TB Memory
Total 8x 3.5" Hot-swap drive bays

2 x 2200W PSU

X86 서버

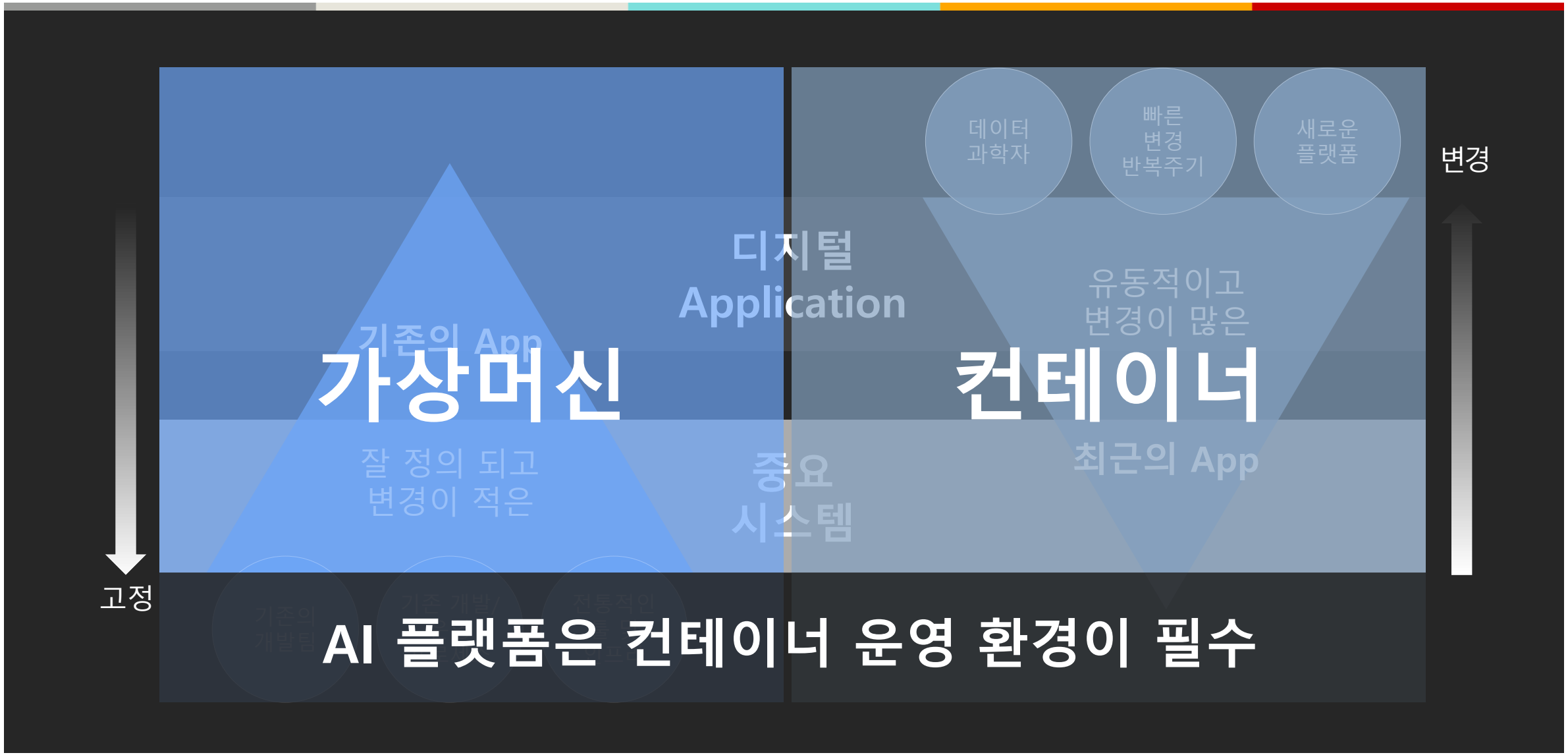
SYS-620P-TR

x86 (2U)

2 x Intel CPU
Max 4TB Memory
8 hotswap 3.5" SATA3 (6 Gbps) drive bays
4 PCI-E 4.0 x16 LP
2 PCI-E 4.0 x8 LP

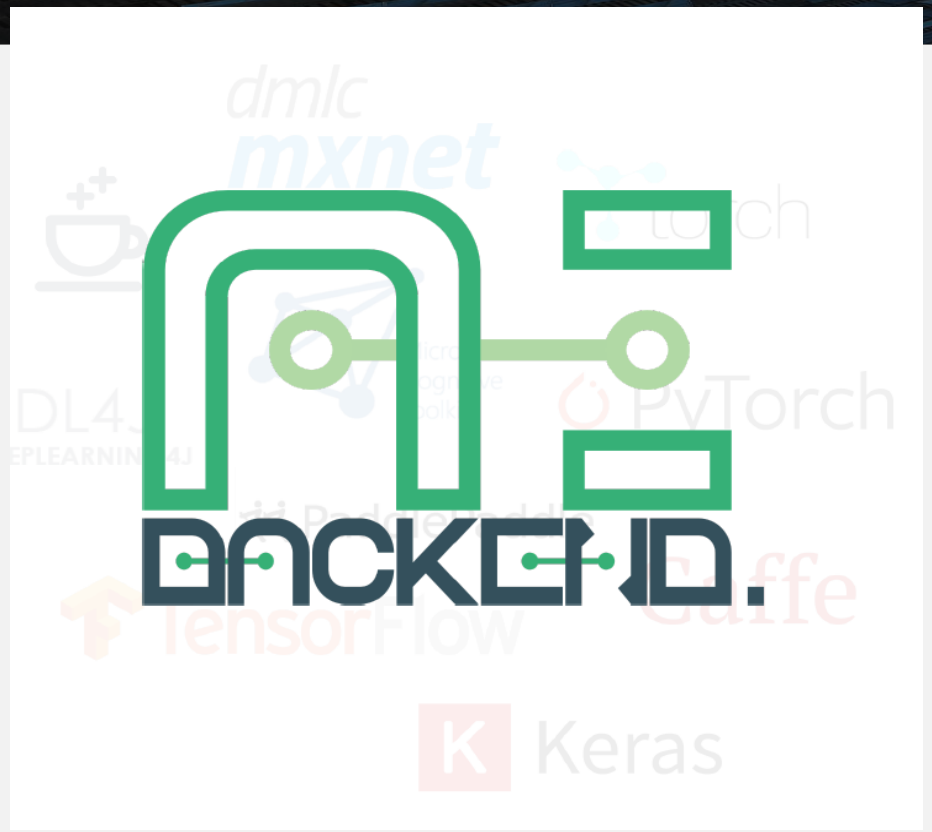
2 x 1200W PSU

2. AI모델 운영



2. AI모델 운영

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



1 GPU 활용 극대화

✓컨테이너수준 GPU분할 가상화
✓NVIDIA GPU MIG 지원

2 직관적인 관리 및 사용자 경험

✓GUI 기반 컨테이너 운영관리
✓웹UI와 데스크탑 앱 지원

3 사전정의 AI개발환경 제공

✓Tensorflow, Pytorch 등 사전정의 이미지 제공
✓연구환경 선택 즉시 생성

4 AI 및 HPC 성능 최적화

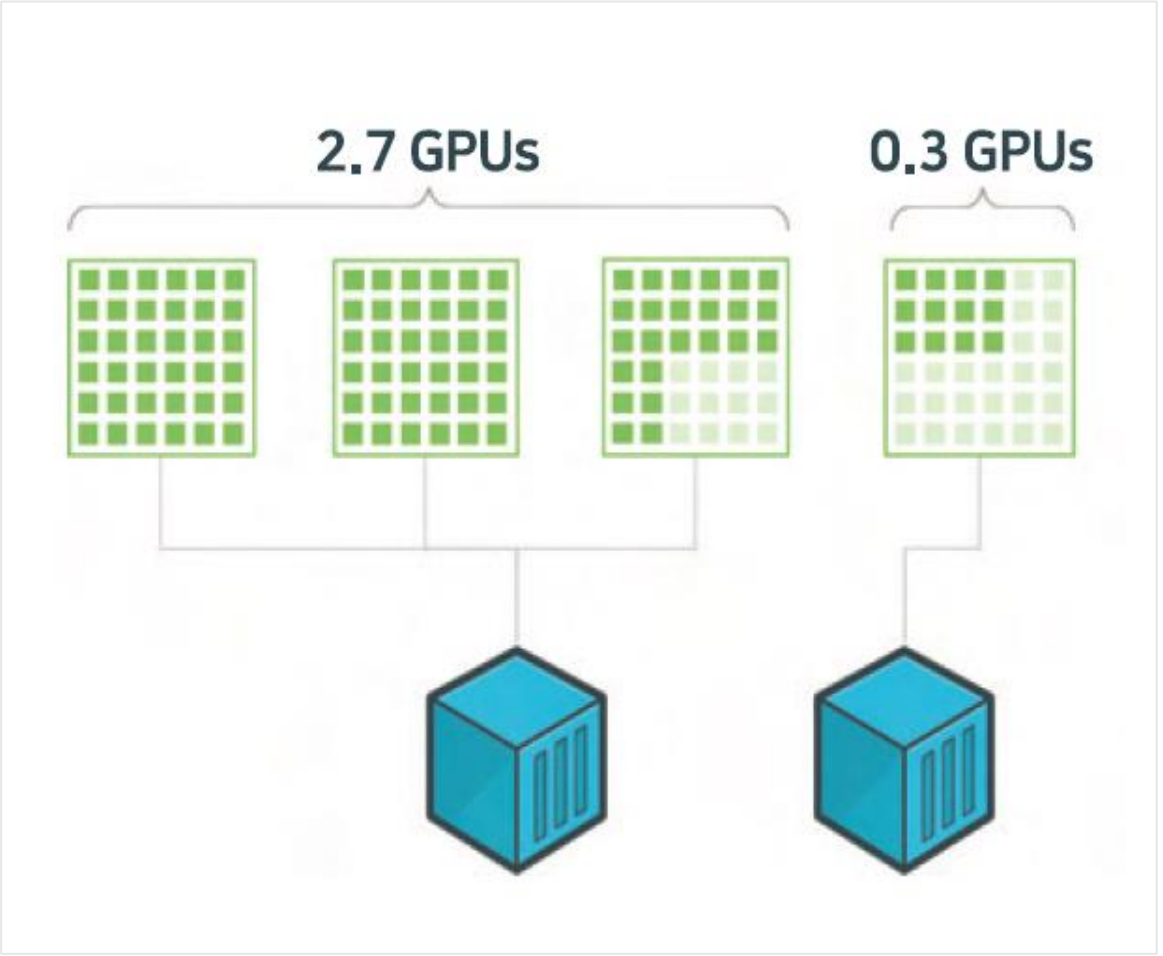
✓독자적 엔진으로 최적의 GPU 연산 자원 배치 구현
✓다중노드 워크로드 및 데이터 I/O 병렬화 지원

Lablup Backend.AI 의 GPU 분할 가상화

컨테이너 기반 GPU 스케일링

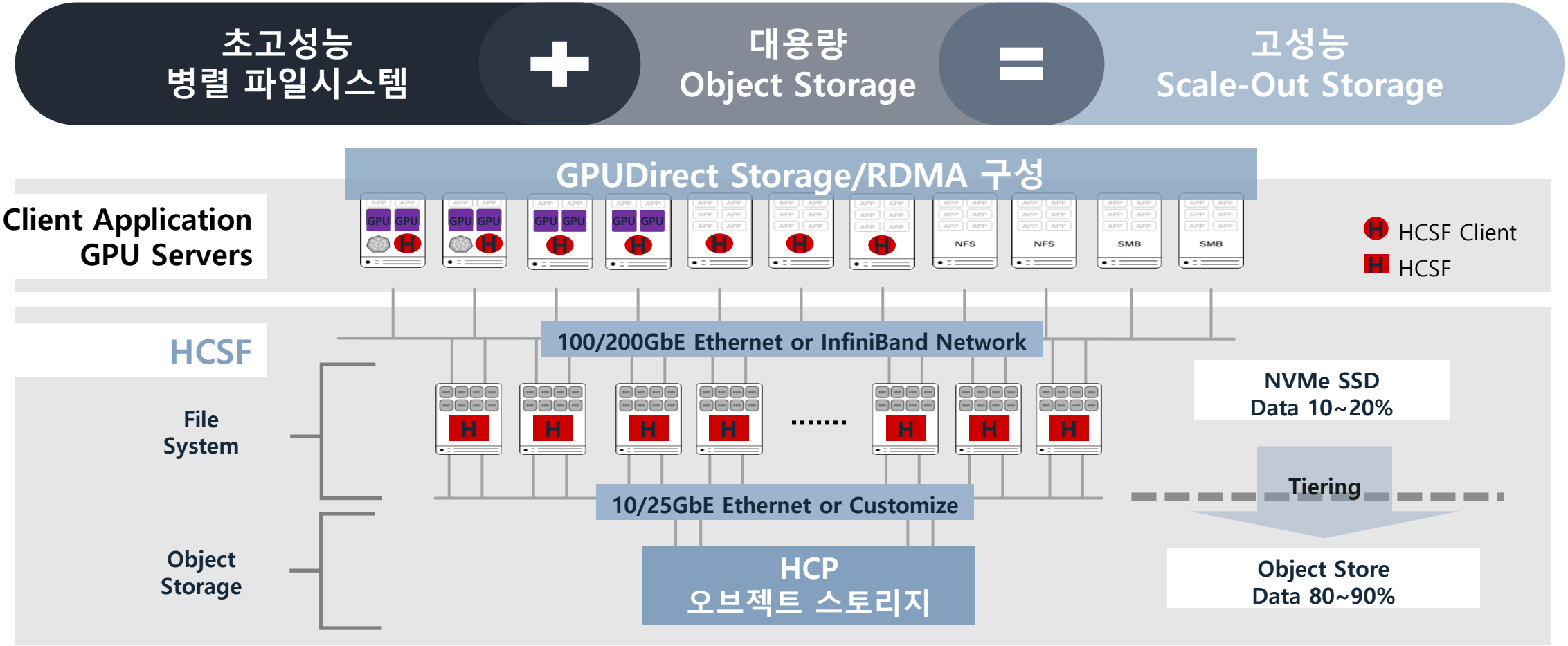
- ✓ 교육 및 추론 워크로드를 위한 단일 GPU 공유
- ✓ 모델 학습 등 대규모 워크로드를 위한 다중 GPU 할당
- ✓ 자체 개발한 CUDA 가상화 계층으로 구현

[Lablup Backend.AI 대한민국, 미국, 일본 등록 특허]



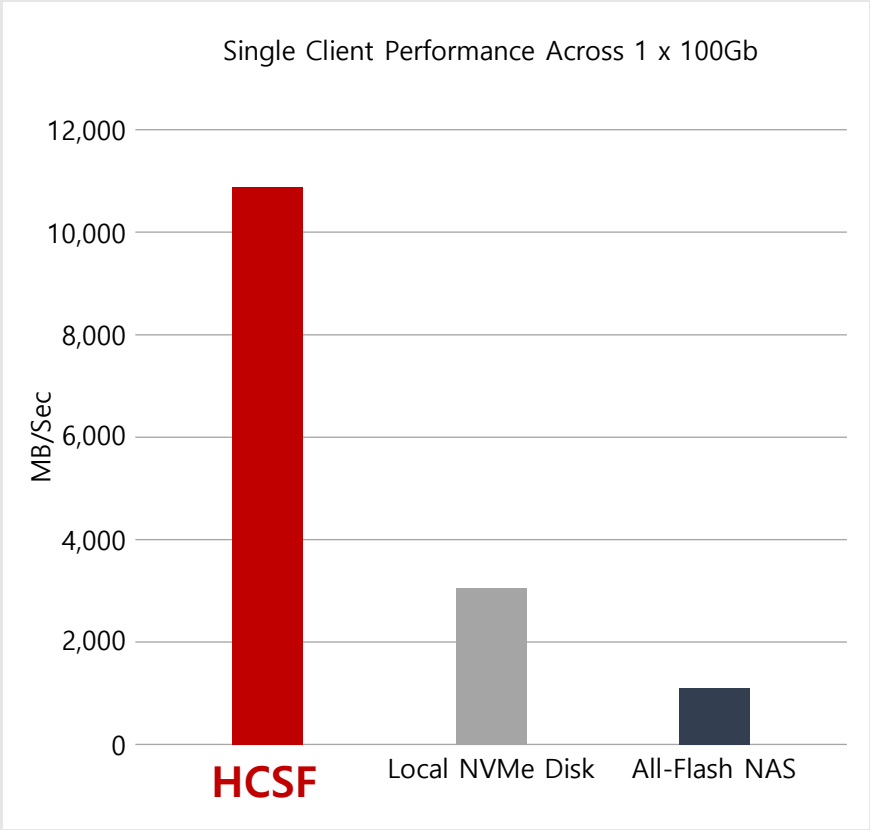
3. 데이터 운영 비용 (HCSF)

효성의 HCSF는 NVMe 기반 Weka IO를 통한 **성능**과 오브젝트 스토리지 Auto Tiering기반 데이터의 **경제적 저장으로 비용 절감**이 가능

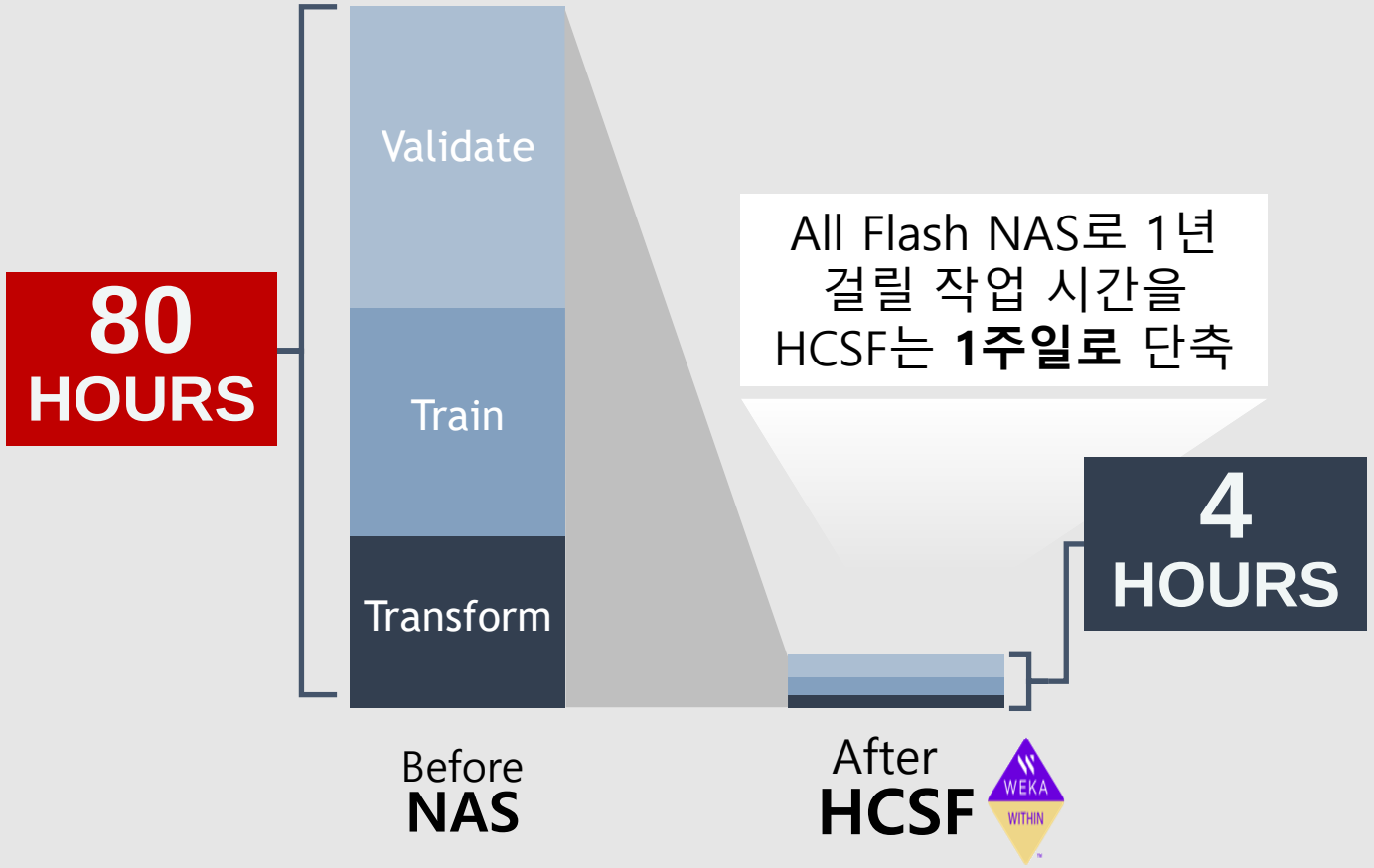


3. 데이터 운영 비용 (최고 성능을 경제적으로 운영)

저장자원 성능 비교 (단일 GPU)



Global T사 사례

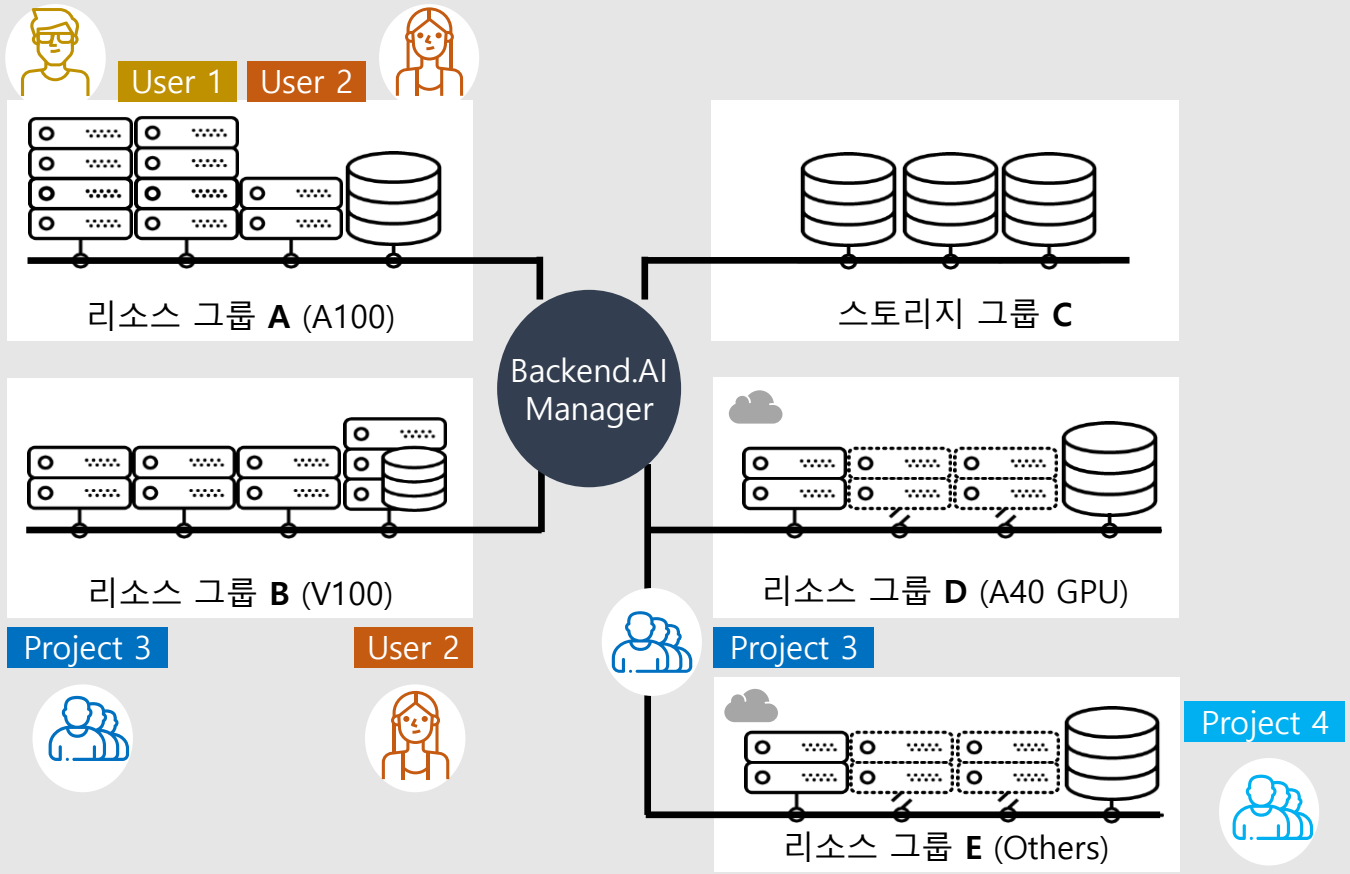


2. 기존 도입 고객 – 기 자원 활용 신규 AI플랫폼 도입

다수의 GPU서버 리소스를 **GPU 종류별** 그룹으로 묶어서 **사용자 및 프로젝트 그룹에 할당**

리소스 그룹 응용 예

- 리소스 그룹 A: NVIDIA A100 GPU 그룹
- 리소스 그룹 B: NVIDIA V100 GPU 그룹
- 사용자 별로 사용 자원을 분리
  - User 1은 A100만을 사용
  - User 2는 A100 및 V100을 모두 사용
- 프로젝트에 따라 사용 자원을 분리
  - 프로젝트 3은 V100 그룹 및 A40 GPU그룹을 사용할 수 있음
  - 프로젝트 4는 Others만 사용



3. 대규모 사업 추진 고객 – 성능 최적화 AI 통합솔루션

운영

AI/ML 모델 개발/운영 효율화
(AI/ML Ops 및 컨테이너 – Lablup Backend.AI)

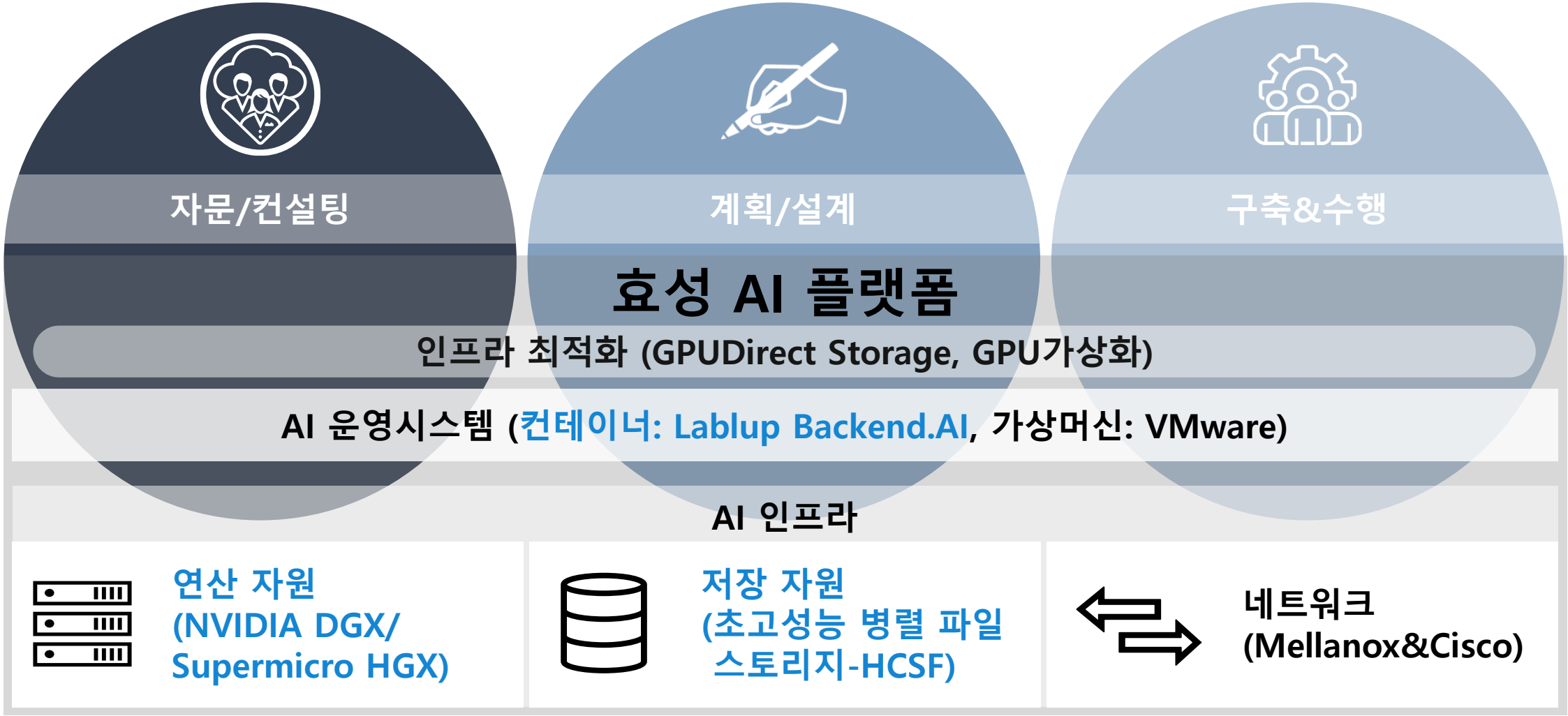
비용

GPU 및 저장 자원의 효율적 사용
(GPU가상화 – 컨테이너 기반 GPU 자원 분할 및 병합 가상화)

성능

스토리지& 네트워크 I/O 성능 최적화
(GPUDirect Storage, 고성능병렬파일 시스템-HCSF)

AI의 시작과 끝 효성이 함께 합니다.




AI의 시작과 끝 효성이 함께 합니다.
- 효성인포메이션시스템 -

Thank you